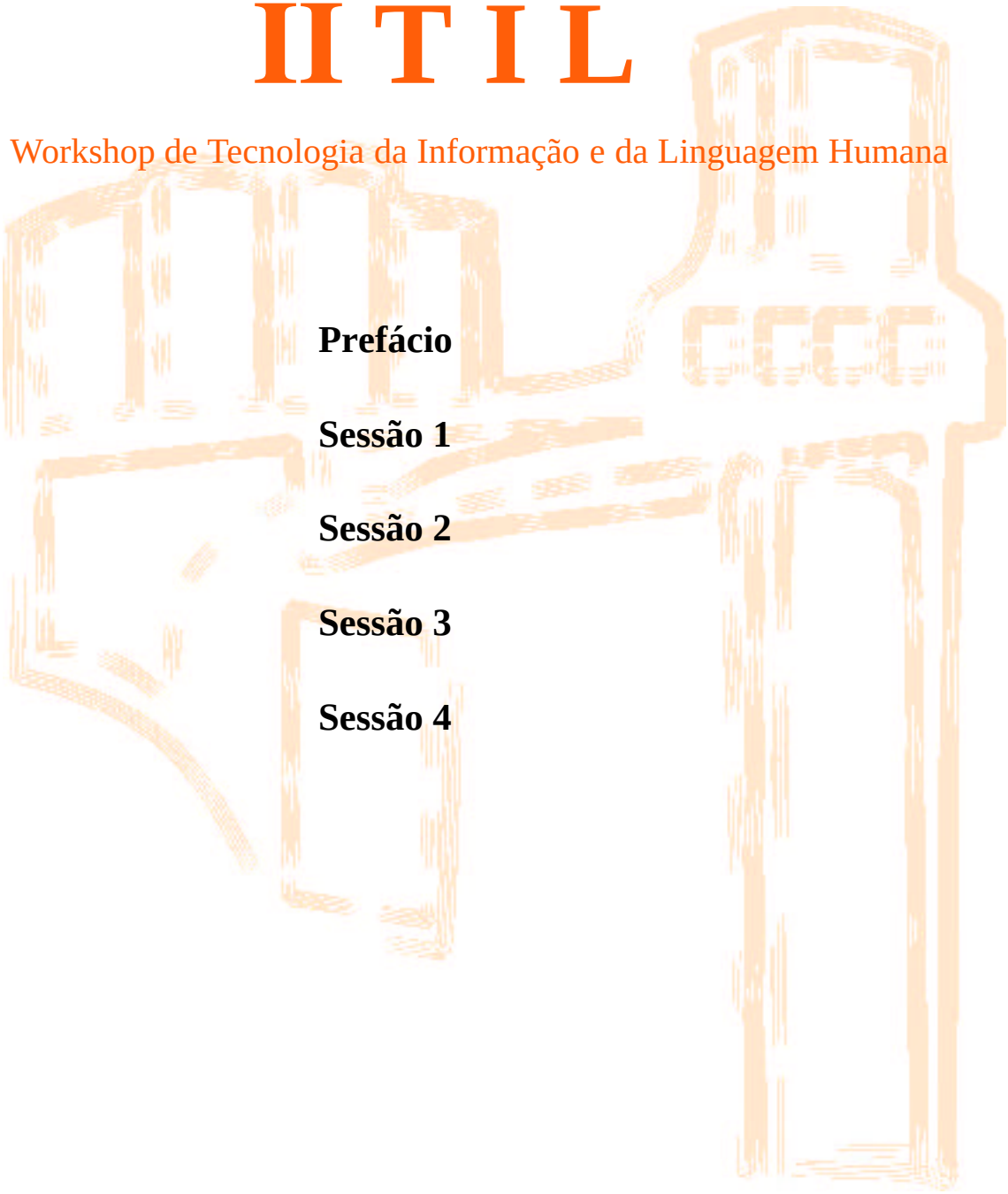
I I T I L

Workshop de Tecnologia da Informação e da Linguagem Humana



II T I L

Workshop de Tecnologia da Informação e da Linguagem Humana

Sessão 1

- Classificação Automática de Textos usando Subespaços Aleatórios e Conjunto de Classificadores
- Uma Arquitetura de Agentes Cooperativos de Informação para a Web Baseada em Ontologias
- Investigação sobre a Identificação de Assuntos em Mensagens de Chat
- Geração de Impressão Digital para Recuperação de Documentos Similares na Web
- Proposta de uma Plataforma para Extração e Sumarização Automática de Informações em Ambiente Web

II T I L

Workshop de Tecnologia da Informação e da Linguagem Humana

Sessão 2

- Abducing Denite Descriptions Links
  - HERMETO: A NL Analysis Environment
  - Impressões Lingüísticas Sobre Duas Axiomatizações para a Gramática Categorial
- Modelos de Linguagem N-grama para Reconhecimento de Voz com Grande Vocabulário

II T I L

Workshop de Tecnologia da Informação e da Linguagem Humana

Sessão 3

- Os Tipos de Anotações, a Codificação, e as Interfaces do Projeto Lácio-Web: Quão Longe Estamos dos Padrões Internacionais para Córpus?
  - Um Modelo de Identificação e Desambigüização de Palavras e Contextos
  - Identificação de Expressões Anafóricas e Não Anafóricas com Base na Estrutura do Sintagma
  - Edição de Informações Sintático-Semânticas dos Adjetivos na Base da Rede Wordnet.Br
  - Locution or Collocation: Comparing Linguistic and Statistical Methods for Recognising Complex Prepositions
- O Problema da Ambigüidade Lexical de Sentido na Comunicação Multilingüe

II T I L

Workshop de Tecnologia da Informação e da Linguagem Humana

Sessão 4

- Identificação do Perfil dos Usuários da Biblioteca Central da FURB Através de Data Mining para a Personalização da Recuperação e Disseminação de Informações

- A Declarative Approach for Information Visualization

- Um Projeto de Metodologia para Escolha Automática de Descritores para Textos Digitalizados Utilizando Sintagmas Nominais

II Workshop de Tecnologia da Informação e da Linguagem Humana
TIL 2004

O TIL 2004, em sua segunda edição, ocorre junto ao Congresso da Sociedade Brasileira de Computação. O Workshop tem por objetivo conjugar propostas de modelagem e manipulação computacional das línguas naturais, ao fomentar a interação entre pesquisadores de várias áreas correlatas à Tecnologia da Informação:

Ciência da Computação

Mineração de Textos na Web ou não, Web Semântica, Recuperação da Informação, Interação Humano-Computador, Banco de Dados Inteligentes, Processamento de Língua Natural Escrita ou Falada, etc., as quais freqüentemente requerem recursos e ferramentas lingüísticas para o projeto e desenvolvimento de sistemas.

Lingüística e/ou Letras

Terminologia, Lexicologia, Construção de Léxicos Semânticos,
Lexicografia, Gramáticas, Análise do Discurso, Construção de Ontologias, Tradução, Lingüística de Corpus, Construção de Dicionários, Modelagem de Ontologias, etc., as quais têm a língua natural como objeto de estudo e a informática, muitas vezes, como instrumento de validação de suas teorias.

Ciência da Informação

Filtragem de Dados, Recuperação da Informação, Catalogação, etc., as quais usam recursos ou modelos de busca de informações relevantes e compatíveis e, muitas vezes, coincidentes com aqueles utilizados, p.ex., no processamento das línguas naturais.

Outras áreas afins, como a de Filosofia ou Ciências Humanas, de um modo geral.

Ao reunir pesquisadores dessas áreas, o Workshop visa ampliar os estudos em sua interface, motivando-os para o conhecimento mútuo de suas pesquisas e respectivas comunidades.
Adicionalmente, o Workshop se apresenta como uma oportunidade de impulsionar as pesquisas envolvendo o português do Brasil, visando, sobretudo, o processamento automático da informação veiculada nessa língua.

A sua realização segue um primeiro encontro realizado em São Carlos em Outubro de 2003, que contou com a participação de 28 instituições, sendo que 21 delas contribuíram com apresentações de trabalhos.
Em 2004 o Workshop recebeu 39 submissões, oriundas de todas as regiões do país e de Portugal, sendo que 18 trabalhos (46%) foram selecionados para apresentação oral e publicação em anais.

Comitê de Programa

Alceu de Souza Britto Jr (PUCPR)
Ariadne Carvalho (UNICAMP)
Bento Carlos Dias da Silva (UNESP)
Carlos Augusto Prolo (PUCRS)
Celso Antônio Kaestner (PUCPR)(Presidente)
Flávio Miguel Varejão (UFES)
Helio Kuramoto (IBICT)
Heronides Moura (UFSC)
José Palazzo Moreira de Oliveira (UFRGS)
Leonardo Lazarte (UNB)
Lígia Café (IBICT)
Lidia Alvarenga (UFMG)
Marco Rocha (UFSC)
Marcos Goldnadel (UNISINOS)
Maria Carmelita P. Dias (PUC-Rio)
Maria Carolina Monard (USP)
Maria das Graças Volpe Nunes (USP)
Marisa Brascher (IBICT)
Oto Araújo Vale (UFGo)
Renata Vieira (UNISINOS)
Rove Chishman (UNISINOS)
Simone Junqueira (PUC-Rio)
Solange Oliveira Rezende (USP)
Stanley Loh (UCPEL)
Vera Lúcia Strube de Lima (PUC-RS)
Violeta Quental (PUC-Rio)

Revisores adicionais
Catia de Azevedo Fronza (UNISINOS)
Julio César Nievola (PUCPR)
Flavio Bortolozzi (PUCPR)

Coordenação Geral do Congresso da SBC2004

Raimundo José de Araújo Macêdo (LaSiD/DCC/UFBA)

Organização

Celso Antônio Alves Kaestner (PUCPR)
Renata Vieira (UNISINOS)

Classificação Automática de Textos usando Subespaços Aleatórios e Conjunto de Classificadores

Chu Chia Gean **Celso Antônio Alves Kaestner**

Programa de Pós-Graduação em Informática Aplicada (PPGIA)
Pontifícia Universidade Católica do Paraná (PUCPR)
Rua Imaculada Conceição, 1155 – 80.215-901 – Curitiba – PR – BRASIL

{ccg,kaestner}@ppgia.pucpr.br

Resumo. *Devido à grande quantidade de informação disponível atualmente em meio eletrônico, a tarefa de classificação automática de textos tem ganhado importância nas pesquisas realizadas na área de Recuperação de Informações. Neste artigo é descrita uma nova abordagem para o problema, fundamentada no modelo vetorial para o tratamento de documentos e em técnicas de reconhecimento de padrões.Como as coleções de textos produzem espaços vetoriais de dimensão elevada, o problema foi atacado pelo uso de diversos procedimentos de pré-processamento e por um conjunto de classificadores k-NN (k vizinhos mais próximos), cada um dos quais dedicado a um subespaço do espaço original. A classificação final é obtida pela combinação dos resultados individuais produzidos por cada classificador. Esta abordagem foi aplicada a coleções de documentos extraídas das bases TIPSTER e REUTERS, e os resultados obtidos são apresentados.*

Abstract. *Nowadays, due to the large volume of text available in electronic media, the automatic document classification becomes an important modern Information Retrieval task. In this paper we describe a new approach to the problem, based on the classical vector space model for text treatment and on a Pattern Recognition approach. As texts collections produce huge dimensional vector spaces, we attack the problem using several preprocessing techniques, and a set of k-Nearest-Neighbors classifiers, each of them dedicated to a sub-space of the original space. The final classification is obtained by a combination of the results of the individual classifiers. We apply our approach to a collection of documents extracted from the TIPSTER and REUTERS databases, and the obtained results are presented.*

1. Introdução

Definitivamente vivemos na era da explosão da informação. Estudos recentes divulgados pela Universidade de Berkeley [Lyman 03] indicam que em 2002 foram criados cerca de 5 milhões de *terabytes* de informação em filmes, em meio impresso, ou em meio de armazenamento magnético ou ótico. Este total é equivalente ao dobro do produzido em 1999, o que indica uma taxa de crescimento da ordem de 30 % ao ano. Somente a WWW agrega em torno de 170 *terabytes*, o que equivale a 17 vezes o tamanho das obras impressas da Biblioteca do Congresso dos EUA.

Por outro lado, o uso das informações disponíveis é muito difícil. Diversos problemas tais como a busca de fontes de informação, a recuperação e extração de informações e a classificação automática de textos tornaram-se importantes tópicos de pesquisa em Computação. O uso de ferramentas automáticas para o tratamento de informações tornou-se essencial ao usuário comum; sem eles se torna praticamente impossível desfrutar de todo o potencial informativo disponível na WWW [Zhong 02].

Em particular, a tarefa de classificação automática de documentos reveste-se de importância, visto que é empregada em diversas tarefas cotidianas, tais como a distribuição e seleção automática de *emails* e a classificação de documentos legados [Belkin 92], [Dhillon 01].

Neste artigo propõe-se uma nova abordagem para o problema da classificação automática de textos, com o uso de subespaços vetoriais do espaço original que relaciona termos e documentos, e de classificadores baseados em instâncias (*k*-vizinhos mais próximos) [Mitchell 97] aplicados a estes subespaços. A classificação final é obtida pela combinação dos resultados individuais dos classificadores aplicados aos subespaços.

O enfoque é testado com o auxílio de duas coleções de documentos largamente empregadas para avaliação da tarefa de classificação automática de textos: a base TIPSTER [Trec 04] e a base REUTERS-21578 [Lewis 04].

O restante deste trabalho é organizado da seguinte forma: a seção 2 apresenta uma visão geral do modelo vetorial utilizado para a representação de documentos, e descreve os procedimentos de pré-processamento e a tarefa objetivo. Na seção 3 é apresentado o formalismo subjacente à proposta. A seção 4 descreve a metodologia empregada para a realização dos experimentos e apresenta dos resultados obtidos. Finalmente a seção 5 apresenta algumas conclusões, perspectivas de trabalho e pesquisas futuras.

2. O modelo vetorial e a classificação automática de documentos

No contexto do tratamento de documentos objetivo principal de um modelo de representação é a obtenção de uma descrição adequada da semântica do texto, de uma forma que permita a execução correta da tarefa alvo, de acordo com as necessidades do usuário.

Diversos modelos têm sido propostos, tais como o modelo booleano [Wartik 92], o modelo probabilista [vanRijsberger 92] e o modelo vetorial [Salton 97]. Neste trabalho é utilizado o modelo vetorial, conforme proposto por Salton; no modelo a unidade básica do texto é denominada *termo*, e pode corresponder a uma palavra, a um radical (*stem*) ou a uma sub-cadeia (*substring*) originária do texto, conforme o procedimento de pré-processamento que será detalhado adiante.

De acordo com o modelo vetorial cada documento é modelado por um vetor no espaço *m*-dimensional, onde *m* é o número de diferentes termos presentes na coleção. Os valores das coordenadas do vetor que representa o documento estão associados aos termos, e usualmente são obtidos a partir de uma função relacionada à freqüência dos termos no documento e na coleção.

Pré-processamento

Na etapa de pré-processamento os documentos, considerados aqui como sendo texto "puro", livre de qualquer formato, são tratados de maneira a produzir uma representação mais compacta que seja mais adequada à realização da tarefa objetivo [Sparck Jones 97].

Uma etapa de pré-processamento típica inclui:

1) A eliminação de palavras comuns: as palavras comuns (*stop words*) são elementos de texto que não possuem uma semântica significativa; sua presença não agrega nenhuma indicação do conteúdo ou do assunto do texto correspondente. Normalmente as palavras comuns são constituídas de artigos, preposições, verbos auxiliares, etc, tais como "*the*", "*a/an/one*", "*in*" ou "*is*". Após sua eliminação obtém-se uma representação reduzida do texto, ainda em formato livre.

2) A obtenção dos radicais (*stems*): em linguagem natural diversas palavras que designam variações indicando plural, flexões verbais ou variantes são sintaticamente similares entre si. Por exemplo as palavras "*delete*", "*deletes*", "*deleted*" and "*deleting*" tem sua semântica relacionada. O objetivo da obtenção dos radicais é a obtenção de um elemento único – o radical – que permita considerar como um único termo, portanto com uma semântica única, estes elementos de texto. Este passo permite uma redução significativa no número de elementos que compõem o texto.

Outra possibilidade de pré-tratamento é a obtenção da representação em *n-grams* do texto [Cavnar 94]: constitui-se em uma representação alternativa, onde os termos são obtidos diretamente como sub-cadeias de comprimento *n* das palavras que compõem o texto original. Por exemplo, a partir da palavra "*house*" e considerando $n = 4$, obtém-se as seguintes 4-grams: "*_hou*", "*hous*", "*ouse*" e "*use_*", onde "_" é usado para indicar o início ou fim da palavra.

Evidentemente os procedimentos (1) e (2) acima descritos exigem conhecimentos lingüísticos do idioma em que o documento foi escrito. Já o uso de *n-grams* é completamente independente de idioma.

O pré-processamento pode ainda incluir uma filtragem dos elementos restantes do texto, com base na freqüência com que os mesmos aparecem no documento ou na coleção. O objetivo desta filtragem é o de limitar o número de termos a serem considerados.

Após a etapa de pré-processamento os documentos podem ser considerados como vetores em conformidade com o modelo vetorial. Os termos podem corresponder diretamente aos elementos de texto, aos *stems*, ou às *n-grams*. A dimensão do espaço vetorial total de documentos corresponde ao número de termos considerados em toda a coleção.

Formalmente, seja $C = \{d_1, d_2, \ldots, d_N\}$ uma coleção não-ordenada de documentos d_i, com M diferentes termos. Então a representação de um documentos será $d_i = (f_{i1}, f_{i2}, \ldots, f_{im})$ para $i = 1$ até N, onde f_{ij} é uma função de avaliação associada ao termo j no documento i. A função de avaliação (ou "peso") f_{ij} mais comumente

utilizada no modelo vetorial é conhecida como métrica *tf* * *idf* [Salton 97], na qual: $f_{ij} = tf_{ij} \ln(\frac{N}{idf_j})$, onde tf_{ij} é a freqüência do termo *j* no documento *i* (*term frequency – tf*), idf_j é o número de documentos que contem o termo *j* na coleção (*inter document frequency – idf*), e *N* é o tamanho da coleção (seu número de documentos). Outras medidas, como a freqüência simples (tf_{ij}), também são usadas (ver [Salton 97]).

Portanto, em conformidade com o modelo vetorial uma coleção de documentos pode ser vista como uma imensa matriz $C_{N \times M}$, onde f_{ij} representa o peso do termo *j* no documento *i*, *M* é o número de termos e *N* é o número de documentos na coleção [Berry 99].

$$C = \begin{bmatrix} f_{11}, f_{12}, \ldots, f_{1M} \\ f_{21}, f_{22}, \ldots, f_{2M} \\ \ldots\ldots\ldots\ldots \\ f_{N1}, f_{N2}, \ldots, f_{NM} \end{bmatrix}$$

Classificação de documentos e o classificador *k*-NN

A classificação de documentos pode ser definida sobre o modelo vetorial como um caso especial de um problema de classificação supervisionada no contexto do Reconhecimento de Padrões [Duda 00].

Considera-se que a coleção de documentos tem uma partição implícita. Cada elemento na partição pertence a uma *classe*, formada pelo subconjunto de documentos que compartilham características comuns. Portanto, pode-se considerar a classe como um atributo especial de cada documento. Um *classificador* é um procedimento que determina, a partir de um documento dado, a sua classe.

Um classificador bem conhecido na área do Reconhecimento de Padrões é o *k*-vizinhos mais próximos (*k*-NN) [Duda 00]. Este algoritmo é amplamente utilizado devido à sua simplicidade conceitual e erro conceitualmente limitado. De maneira abreviada um classificador *k*-NN associa a um documento *d* à classe mais freqüente entre as classes dos *k* vizinhos mais próximos de *d* na coleção, de acordo com uma distância calculada no espaço vetorial de documentos.

Na área do tratamento de textos as distâncias entre dois documentos d_i e d_j mais comumente utilizadas são a distância euclidiana $dist(d_i, d_j) = \left[\sum_{k=1}^{M} (f_{ik} - f_{jk})^2\right]^{1/2}$ e a denominada "métrica do co-seno" $cos(d_i, d_j) = \frac{d_i * d_j}{\|d_i\| * \|d_j\|}$ [Salton 97].

3. Subespaços aleatórios e combinação de classificadores

Devido à dimensão elevada do espaço de documentos (*M*), propõe-se neste trabalho a divisão do espaço original em diversos subespaços, cada qual tratado por um classificador específico.

Considere-se o caso de *P* subespaços: inicialmente algumas colunas da matriz de (documentos x termos) *C* são selecionadas aleatoriamente. Se *1,2,…M* são as colunas de

C, seja *X* o subespaço projeção sobre estas colunas; *proj $_X$ (C)* representa a sub-matriz obtida de *C* pela projeção de suas linhas sobre *X*, com dimensão *N* x | *X* |, e *proj $_X$ (d)* é a matriz 1 x | *X* | que corresponde a um documento *d*.

Em cada subespaço gerado desta forma um classificador pode atuar. Nos experimentos constantes deste trabalho foram utilizados subespaços de mesma dimensão (isto é | *X* | é constante para cada subespaço *X*). Em cada *X* empregou-se um classificador *k*-NN fundamentado na métrica do co-seno com o critério usual de classificação do algoritmo. Por exemplo, para *k*=1 segue-se o seguinte critério de classificação: Classe (*d*) = Classe(d_i) onde d_i é tal que $cos(d_i,d) < cos(d_j,q)$ para todo $j <> i$.

Quando se aplica a regra de classificação em cada subespaço, obtem-se *P* possivelmente diferentes classificações. Então se deve decidir a classe de *d* usando um procedimento de decisão que leve em conta os resultados individuais dos diferentes classificadores de 1 até *P*. Usualmente para a combinação de classificadores se emprega o princípio do voto da maioria (*majority vote principle*), isto é, assinala-se ao documento *d* a classe mais freqüente entre as *P* assinaladas individualmente pelos classificadores a *d*.

Além desta regras, neste trabalho empregou-se uma segunda regra de combinação: inicialmente um conjunto com todos os documentos que se constituem nos vizinhos mais próximos a *d* é formado; em seguida determina-se a classe de cada um destes documentos e a mais freqüente é indicada. Este procedimento considera apenas documentos diferentes para calcular a classe final, visto que a formação do conjunto intermediário elimina aparecimentos múltiplos dos documentos, não importando o número de vezes em que os mesmo apareçam nas *P* classificações.

O método delineado acima, com o uso de subespaços vetoriais do espaço original de características e o emprego de combinação de classificadores é uma variante da discriminação estocástica, onde diversos classificadores criados estocasticamente são combinados de forma a aumentar a correção preditiva. Este método tem sido utilizado com sucesso em outros domínios, como por exemplo, no reconhecimento de imagens de dígitos manuscritos [Ho 98].

4. Experimentos realizados e resultados obtidos

Para verificar a aplicabilidade dessa abordagem para a classificação automática de documentos, alguns experimentos preliminares já foram realizados e são descritos a seguir neste trabalho.

Os testes foram realizados utilizando-se duas coleções: (1) a coleção TIPSTER, da conferência TREC [Trec 04], uma competição para a avaliação de sistemas de tratamento automático de documentos; e (2) a coleção REUTERS-21578 [Lewis 04], que foi especificamente construída para a avaliação de sistemas de classificação e é largamente utilizada na literatura da área.

A coleção TIPSTER é formada por milhares de documentos em Inglês (em formato XML), com tamanhos variando de uma a duas linhas até uma ou duas páginas. Os documentos estão agrupados em séries formadas por milhares de elementos. A TREC não possui uma tarefa específica de classificação de documentos; no entanto a

partir da tarefa de recuperação de documentos – quando a partir de uma consulta do usuário deve ser recuperada uma lista ordenada de documentos relevantes – é possível se obter uma partição da coleção em classes: são considerados similares documentos que responder a uma mesma consulta. A indicação da relevância dos documentos em relação às consultas foi feita manualmente por um grupo de especialistas.

Para se obter uma coleção adequada à tarefa de classificação foram selecionados, para experimentos preliminares, 60 documentos que são considerados relevantes para 5 consultas, formando uma coleção equilibrada de 5 classes com 12 elementos cada.

No primeiro experimento os documentos foram pré-processados usando-se a eliminação de palavras comuns e a obtenção dos radicais. A lista de palavras comuns que foram eliminadas foi obtida da BOW Library – CMU e utilizou-se o algoritmo de Porter [Porter 97] para o procedimento de *stemming*. No total foram produzidos 2611 termos, gerando uma matriz $C_{60 \text{ X } 2611}$; os elementos de C foram calculados usando a freqüência simples, isto é, com $f_{ij} = tf_{ij}$.

Dos documentos da base 45 foram utilizados para treinamento e 15 para teste. Foram empregados 30 subespaços aleatórios ($P = 30$), cada um dos quais com dimensão 50 ($| X | = 50$). Em cada subespaço empregou-se um classificador k-NN de funcionamento padrão, usando a métrica do co-seno como medida de similaridade. A combinação dos resultados dos classificadores aplicados aos subespaços foi feita de acordo com as duas regras de combinação já descritas: (1) na primeira delas Classe (d) é a classe mais freqüente retornada pelos classificadores; e (2) Classe(d) é obtida como a classe mais freqüente entre os documentos que constituem os k vizinhos retornados por cada classificador, anteriormente agrupados em um único conjunto.

Os resultados obtidos são sumarizados à Tabela 1, em função dos diferentes valores do parâmetro k. A medida empregada para a avaliação é a *correção*, definida como a porcentagem dos documentos corretamente classificados.

Tabela 1: Correção (em %) segundo os diferentes parâmetros, 1º experimento

k	1ª regra para combinação (*majority vote*)	2ª regra para combinação das classificações
1	50,0	93,3
2	66,7	66,7
3	66,7	60,0

Pode-se observar que, surpreendentemente, os melhores resultados foram obtidos para $k = 1$, e que a segunda regar de combinação de classificadores produz resultados superiores.

No segundo experimento os documentos foram pré-processados utilizando-se a eliminação de palavras comuns e aplicação posterior do processo de obtenção de 4-*grams*. Usou-se a mesma lista de *stop-words* (BOW Library) e um procedimento padrão para obter as 4-*grams* [Cavnar 94]. No total foram produzidos 7027 termos, gerando uma matriz $C_{60 \text{ X } 7027}$, cujos elementos foram obtidos por freqüência simples, como no primeiro experimento. A partição utilizada para treinamento e testes (75 % e 25 %) foi a mesma; também se utilizaram 30 subespaços aleatórios ($P = 30$). Para levar em conta a

maior dimensionalidade do espaço produzido pelas 4-*grams*, empregaram-se subespaços de dimensão 150 ($| X | = 150$). Os classificadores utilizados também foram idênticos aos do primeiro experimento: *k*-NN com uso da métrica de similaridade do co-seno.

Os resultados obtidos são sumarizados à Tabela 2, usando a mesma unidade de avaliação: a taxa de correção na classificação.

Tabela 2: Correção (em %) segundo os diferentes parâmetros, 2º experimento

k	1ª regra para combinação (*majority vote*)	2ª regra para combinação das classificações
1	53,3	66,7
2	53,3	53,3
3	53,3	60,0

Estes resultados são compatíveis com os obtidos no primeiro experimento: aqui novamente a segunda regra de decisão produz resultados superiores.

Em seguida foram realizados experimentos utilizando-se a coleção de documentos REUTERS-21578 [Lewis 04]. Esta base é formada por documentos em XML, permitindo que se indique no corpo do documento as classes ao que o mesmo pertence, segundo diversas classificações. As categorias disponíveis são, por exemplo, <Date>; <Topic>; <Place>; <People>; <Orgs>; <Exchanges>; etc. .

Nos experimentos realizados utilizaram-se somente os 1000 documentos que constituem o primeiro grupo da base em questão, e uma única categoria (<Place>) para a determinação das classes. Neste grupo esta categoria constitui 133 classes, das quais as mais freqüentes são "USA" com freqüência 474, a ausência de informação – que aparece 150 vezes; e a classe "UK", com 50 exemplos. Por outro lado 89 classes possuem um único exemplo neste grupo.

O pré-processamento constitui-se da eliminação de palavras comuns, obtenção de radicais, e exigência do aparecimento do termo em no mínimo dois documentos. Obteve-se assim 3633 termos e conseqüentemente uma matriz $C_{1000 \times 3633}$.

A partição utilizada para treinamento e testes foi de 70 % e 30 %, respectivamente. Foram utilizados 30 subespaços vetoriais (P=30) de dimensão $|X|$ = 1000 cada. Foram efetuados experimentos com a função de ponderação $f_{ij} = tf_{ij}$ (freqüência simples) e também com: $f_{ij} = tf_{ij}\ idf_j$ (métrica *tf*idf*). Os resultados obtidos em termos da taxa de correção são apresentados à Tabela 3.

Tabela 3: Correção (em %) segundo os diferentes parâmetros, 3º experimento

f_{ij}	*k*	1ª regra para combinação (*majority vote*)	2ª regra para combinação das classificações
tf	1	59,7	60,3
tf	2	59,7	59,7
*tf*idf*	1	64,7	63,3
*tf*idf*	2	63,0	60,0

Os resultados obtidos preliminarmente nestes três experimentos são compatíveis com outros experimentos relatados na literatura realizados em condições semelhantes, e podem ser considerados como aceitáveis em diversas aplicações práticas de classificação automática ou semi-automática de documentos.

5. Conclusões e trabalho futuros

Este artigo apresenta uma nova proposta para a realização da tarefa de classificação automática de documentos por meio do uso de subespaços vetoriais do espaço original que relaciona termos e documentos.

Neste trabalho utiliza-se o modelo vetorial para a representação de documentos, de forma que a aplicação da proposta é direta. São empregados conjuntos de classificadores *k* vizinhos mais próximos (*k*-NN) e regras para a combinação dos resultados obtidos individualmente por cada classificador.

Os resultados obtidos, embora preliminares, são encorajadores e indicam a aplicabilidade do método.

Está prevista a realização de novos experimentos para uma melhor avaliação da proposta, nas seguintes direções:

1) Aplicação da proposta a uma coleção de maior envergadura, apara avaliar sua escalabilidade;
2) Avaliação mais detalhada dos efeitos do pré-processamento, incorporando outras combinações relacionadas à eliminação de palavras comuns, obtenção de radicais, obtenção de *n-grams*, e de outros filtros;
3) Realização de testes para avaliar a sensibilidade da arquitetura proposta em relação aos diferentes parâmetros envolvidos, tais como a dimensão do subespaço ($|X|$), e variações no número (P) e no tipo dos classificadores, com uso de árvores de decisão, Naïve-Bayes, e outros algoritmos de classificação [Deb 01], [Mitchell 97]; e
4) Uso de técnicas mais sofisticadas para a seleção dos subespaços a considerar, como o emprego da Análise Semântica Latente (LSA) e suas variações [Deerwester 90], [Zha 98], [Zha 98b].

6. Referências

[Baeza-Yates 99] Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley, 1999.

[Belkin 92] Belkin, N.; Croft, W. "Information Filtering and Information Retrieval: Two Sides of the Same Coin". *Communications of the ACM*, Nº 35, pp. 29-38, 1992. .

[Berry 99] Berry, M.; Drmac, Z.; Jessup, E. "Matrices, Vector Spaces, and Information Retrieval", *SIAM Review*, Vol. 41, Nº 2, pp.335-362, 1999.

[Cavnar 94] Cavnar, W. B. "Using An N-Gram-Based Document Representation With a Vector Processing Retrieval Model". In *Proceedings Of TREC-3 (Third Text Retrieval Conference)*. Gaithersburg, Maryland, USA, 1994.

[Deb 01] Deb, K. *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons, 2001.

[Deerwester 90] Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T. "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*, Vol. 41, Nº 6, pp. 391-407, 1990.

[Dhillon 01] Dhillon, I.; Modha, D. "Concept Decompositions for Large Sparse Text Data using Clustering". *Machine Learning*, Vol. 42, Nº 1, pp. 143-175, 2001.

[Duda 00] Duda, R.; Hart, P.; Stork, D. *Pattern Classification (2nd. Edition)*, Wiley Interscience, 654 p., 2000.

[Ho 98] Ho, T.K. "The Random Subspace Method for Constructing Decision Forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, Nº 8, pp. 832-844, 1998.

[Lewis 04] Lewis, D.D. http://www.daviddlewis.com/resources/testcollections/reuters21578/; acessado em [08/03/2004].

[Lyman 03] Lyman, P. and Varian H.R. (2003). How Much Information. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 acessado em [19/01/2004].

[Mitchell 97] Mitchell, T. *Machine Learning*. McGraw-Hill, 414p., 1997.

[Porter 97] Porter, M.F. "An algorithm for suffix stripping". *Program 14*, 130-137. 1980. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) *Readings in Information Retrieval*. Morgan Kaufmann, pp. 313-316, 1997.

[Salton 97] Salton, G.; Buckley, C. "Term-weighting approaches in automatic text retrieval". *Information Processing and Management 24,* 513-523. 1988. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) *Readings in Information Retrieval*. Morgan Kaufmann, pp. 323-328, 1997.

[Sparck-Jones 97] Sparck-Jones, K.; Willet, P. (Eds.) *Readings in Information Retrieval*. Morgan Kaufmann, 1997.

[Trec 04] http://trec.nist.gov/data.html; acessado em [08/03/2004].

[van Rijsbergen 92] van Rijsbergen, C.J. Probabilistic retrieval revisited. *The Computer Journal*, Vol. 35, No. 3, pp. 291-298, 1992.

[Wartik 92] Wartik, S. "Boolean Operations". *In Information Retrieval: Data Structures and Algorithms*. Frakes, W.B.; Baeza-Yates, R. (Eds.), Prentice Hall, pp. 264-292, 1992.

[Zha 98] Zha, H.; Simon, H. "On Updating Problems in Latent Semantic Indexing". *SIAM Journal of Scientific Computing*, Vol. 21, pp. 782-791, 1999.

[Zha 98b] Zha, H.; Marques, O.; Simon, H. "A Subspace-Based Model for Information Retrieval with Applications in Latent Semantic Indexing". IRREGULAR '98, Berkeley, California, USA, *Lecturer Notes in Computer Science* Nº 1457, Springer Verlag, pp.29-42, 1998.

[Zhong 02] Zhong, N.; Liu, J.; Yao, Y. "In Search of the Wisdom *Web*". *IEEE Computer*, Vol. 35, Nº 1, pp. 27-31, 2002.

Uma Arquitetura de Agentes Cooperativos de Informação para a Web Baseada em Ontologias

Fred Freitas[1], Tércio de Morais Sampaio[1], Rafael Cobra Teske[2], Guilherme Bittencourt[2]

[1]Departamento de Tecnologia de Informação – Universidade Federal de Alagoas
Campus A.C. Simões - BR 104 - Km 14 - Tabuleiro dos Martins – 57.072-970
Maceió – AL – Brasil

[2]Departamento de Automação e Sistemas - Universidade Federal de Santa Catarina
Caixa Postal 476 - 88.040-900 - Florianópolis - SC – Brasil

fred.freitas@tci.ufal.br,terciomorais@yahoo.com,{cobra,gb}@das.ufsc.br

Abstract. *In the Web, there are classes of pages (e.g., call for papers' and researchers' page), which are interrelated forming clusters (Science). We propose a reusable architecture for multi-agent systems to retrieve and classify pages from these clusters, supported by data extraction. Crucial requirements: (a) a Web vision coupling this vision for contents to a functional vision (role of pages in data presentation); (b) ontologies to represent agents' knowledge about tasks and the cluster. This web vision and agents' cooperation accelerate retrieval. We got quite promising results with two agents for the page classes of scientific events and articles. A comparison with WebKB comes up with a new requirement: a detailed ontology cluster.*

Resumo. *Existem, na Web, classes de páginas (e.g. "call for papers", pesquisadores) que inter-relacionam-se, formando* grupos *(o meio científico). Propomos uma arquitetura reusável de sistemas multiagentes cognitivos para recuperar e classificar páginas destes grupos, baseada na extração de dados. Requisitos: (a) uma visão da Web que acopla a visão por conteúdo a uma visão funcional (papel das páginas na apresentação de dados); (b) ontologias sobre as tarefas dos agentes e o* grupo. *Esta visão da Web e a cooperação entre agentes aceleram a recuperação. Obtivemos bons resultados com dois agentes para as classes de eventos e artigos científicos. Uma comparação com o WebKB sugere um novo requisito: uma ontologia detalhada do* grupo.

1. Introdução

Apesar do termo "agentes cooperativos de informação" ser muito citado, há poucos sistemas na Web que mostram alguma forma de cooperação ou integração entre tarefas relacionadas a texto, como recuperação, classificação e extração de dados. Na definição de manipulação integrada de informação, a extração é mencionada como técnica de aquisição de conhecimento, sugerindo implicitamente integração entre as tarefas. Os sistemas de extração atuam sobre domínios muito restritos, como notícias sobre terrorismo, classificados. Um fato negligenciado sobre as classes de páginas processadas por extratores é que muitas delas inter-relacionam-se, formando *grupos* (*clusters*).

Visando integração e cooperação, projetamos uma arquitetura de sistemas multi-agentes que recupera e classifica páginas de grupos de classes inter-relacionadas na Web, extraindo dados delas. Para permitir a cooperação, dois requisitos são cruciais: (a) uma visão da Web que acopla a visão por conteúdo a uma visão funcional (papel das páginas na apresentação de dados); (b) ontologias sobre as tarefas dos agentes e o *grupo*. São relatados promissores resultados com os agentes de eventos e artigos científicos, com. A categorização funcional, e as listas em particular, melhoram a busca e uma comparação com o WebKB sugere o uso de uma ontologia detalhada do *grupo.*

O artigo está assim organizado: A Seção 2 descreve a visão da Web proposta. A seção 3 introduz a arquitetura e seus componentes. A seção 4 apresenta um estudo de caso, o sistema MASTER-Web (*Multi-Agent System for Text Extraction, classification and Retrieval over the Web*), aplicado ao grupo científico, com dois agentes, um de eventos e outro de artigos científicos. Os resultados nas classificações por conteúdo e funcional são apresentados. A seção 5 compara o MASTER-Web com sistemas similares, e a seção 6 traz trabalhos futuros e conclusões.

2. Visão da Web para a Manipulação Integrada de Informação

As páginas que apresentam entidades compartilham estilo de editoração, , terminologia e o conjunto de atributos. A criação de extratores baseia-se neste fato, na existência de *classes de páginas*, definindo uma visão da Web por conteúdo. Em páginas de pesquisadores, por exemplo, encontram-se dados como instituições, áreas de interesse, artigos e muitos outros itens. O conjunto de atributos de uma classe é *discriminante,* no sentido em que sua presença ajuda a distinguir instâncias da classe [Rilloff 94].

Muitos ponteiros das páginas que pertencem a estas classes apontam para outras páginas contendo entidades ou atributos ou âncoras pertencentes a um número reduzido de outras classes. Chamamos a este conjunto de classes *grupo (cluster) de classe*. Por exemplo, em páginas de pesquisadores, com certeza serão encontrados ponteiros para páginas de artigos, podem ser localizados ponteiros para chamadas de trabalho de eventos científicos, e páginas de outras classes.

Uma visão alternativa da Web diz respeito à funcionalidade das páginas, dividindo-as de acordo com o seu papel na ligação e na apresentação dos dados. Visando a extração integrada, as categorias funcionais se dividem em: *páginas-conteúdo,* (membros de uma classe), *listas* de páginas-conteúdo, *mensagens* (sobre uma classe), *recomendações* (páginas de outras classes), ou *lixo*.

Tanto para identificar precisamente as páginas com instâncias das classes processadas, como para localizá-las rapidamente, beneficiando-se dos relacionamentos entre elas refletidos nas âncoras, as duas visões devem ser usadas simultaneamente.

3. Arquitetura Proposta

Propomos uma arquitetura de Sistemas Multiagentes Cognitivos [Freitas & Bittencourt 2003] para resolver o problema da extração integrada de páginas-conteúdo pertencentes às classes que integram um grupo (*cluster).* A motivação principal para o emprego de sistemas multiagentes é beneficiar-se dos relacionamentos entre as classes. A visão geral da arquitetura está ilustrada na figura 1.

Cada agente, representado como um círculo na figura, reconhece, filtra e classifica páginas que, supostamente, pertençam à classe de páginas processada por eles (por exemplo, páginas de CFPs, artigos e outros para o grupo científico), extraindo também seus atributos. Uma vez que os agentes cooperam, possuindo, porém, responsabilidades distintas, a arquitetura baseia-se na abordagem de Resolução Distribuída de Problemas (RDP) [Álvares & Sichman 95]. A estrela indica troca de mensagens contendo regras de reconhecimento e fatos (conhecimento dos agentes), além das URLs sugeridas entre os agentes.

Cada agente possui um meta-robô, que se conecta a múltiplos mecanismos de busca - como *Altavista, Excite, Infoseek* e outros. Ele consulta os mecanismos de busca com palavras-chave que garantem cobertura em relação à classe de páginas processada pelo agente. (e.g., os termos *'call for papers'* e *'call for participation'* para o agente CFP). Devido à falta de precisão, o conjunto de páginas resultante das consultas recai em vários grupos funcionais além do de páginas-conteúdo tratado, apresentando muitas listas, mensagens, páginas-conteúdo de outras classes, e lixo. As URLs são dispostas numa fila de URLs de baixa prioridade. Cada agente continuamente acessará, além desta fila, outra de alta prioridade, que armazena URLs sugeridas por outros agentes ou presentes em páginas da categoria funcional listas. Afinal, estes endereços são obtidos dentro de um contexto mais confiável e com maior probabilidade de ser relevante do que as listas de resultados dos mecanismos de buscas.

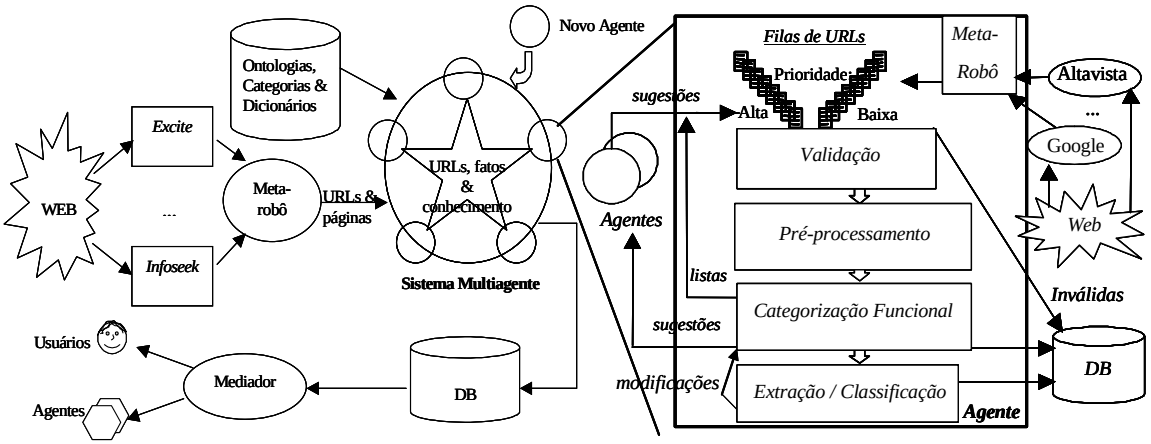


Figura 1. Visão geral da arquitetura de sistemas multiagentes para manipulação integrada, mostrando o funcionamento de um agente em detalhe.

Um mediador estará disponível, com a função de ajudar às consultas aos dados, provendo visões não-normalizadas - mais simples - da base de dados, e permitindo a qualquer usuário ou agente beneficiar-se do acesso aos dados extraídos.

Ao entrar no sistema, os agentes registram-se e anunciam-se aos outros agentes, mandando fatos e regras de reconhecimento de páginas e ponteiros úteis a si próprio, que serão empregadas pelos outros para lhe indicarem sugestões de páginas. O novo agente receberá, reciprocamente, regras úteis aos outros agentes. Assim, quando um agente acha informação que dispara alguma das regras referentes aos outros, este agente repassa a informação (ponteiro ou página) ao agente que lhe enviou a regra disparada.

3.1. Tarefas dos Agentes

Um agente executa quatro tarefas em cada página que processa:

- Validação: Nesta fase são eliminadas páginas inacessíveis, já existentes no BD e em formatos que os agentes não possam processar.

Pré-processamento: Representa as páginas de várias maneiras, tais como conteúdo com e sem HTML, palavras-chave e freqüências, ponteiros, e-mails, e outros, com dados extraídos delas, aplicando, se necessário, recuperação de informação e processamento de linguagem natural (PLN). Os dados passam ao motor de inferência.

- Categorização Funcional: Aqui, as páginas são classificadas em grupos funcionais e são encontradas e enviadas as sugestões para outros agentes, quando uma das regras enviadas por eles dispara. Por exemplo, uma âncora com a palavra "*conference*" é útil para o agente CFP. As sugestões podem, inclusive, acionar buscas em diretórios com prefixo comum, como `/staff/` para o agente de pesquisadores.

- Extração e Classificação: São extraídos os atributos, armazenados na base de dados, ou, pela inconsistência ou inexistência destes, corrigida a classificação de página em relação aos grupos funcionais. Por exemplo, a presença de datas com mais de um ano de intervalo numa página de CFP, denunciam uma página confundida por um CFP.

Lançando mão das representações necessárias, a extração é efetuada por uma combinação de *templates*, e regras ou uma categoria é inferida, normalmente quando achados termos dos dicionários (e.g. a presença de siglas de estados norte-americanos). Após a extração, os atributos podem, adicionalmente, ser formatados (e.g., datas). A partir daí, testam-se casos que contém um conjunto mínimo de atributos para identificar páginas conteúdo. Um exemplo está disposto na próxima seção.

Após a identificação de uma página conteúdo, outros dados são procurados e extraídos. Até este ponto, os dados extraídos apenas evidenciavam a existência de determinados itens de informação. Entretanto, os dados extraídos não são devidamente contextualizados. Por exemplo, datas extraídas de páginas de CFP podem ter significados diversos, como data limite para entrega de trabalhos – *deadline*, data de notificação, data do evento, etc.

Na etapa de extração, combinações de diversos dados extraídos podem compor informações cuja extração é desejada. Para isto, são usadas instâncias de *templates* onde são definidos quais dados se relacionam e como ocorre esta relação de modo a formar uma informação consistente. Um exemplo de extração encontra-se na seção 4.

3.2. Conhecimento dos Agentes

Ontologias desempenham um papel fundamental na arquitetura, servindo não só como vocabulário de comunicação entre agentes, como também na definição e organização apropriadas de conceitos, relações, e restrições. Quatro ontologias se fazem necessárias:

- Ontologia do domínio (ou grupo): Ontologia principal, devendo ser bastante detalhada para garantir precisão à classificação por conteúdo (ver subseção 5.1.).

- Ontologia da Web: Contém definições de *hyperlink*, termo e freqüência, e de página da Web em suas várias representações e atributos - como listas de palavras-chaves e

freqüências, ponteiros, e-mails, etc. Pode, ainda, conter definições relativas à Internet, como protocolos e tipos de arquivos, além de representações de páginas em PLN.

- Ontologia de manipulação integrada de informação: Classes e instâncias empregadas na extração e classificação funcional e por conteúdo. Inclui

– *Templates* reconhecedores das categorias funcionais e páginas-conteúdo

– *Templates* extratores e classificadores de dados

– Classes auxiliares, como meta-definições de conceitos e sinônimos e palavras-chave, classes de PLN (tendo como atributos *parts-of-speech-tags como* rótulos de frases/sintagmas), agentes e habilidades, etc

– Casos complexos que identificam atributos, classes de páginas, categorias funcionais e sugestões para outros agentes. Os casos devem ser bastante expressivos, com conjuntos de atributos e conceitos cuja presença e/ou ausência implica que em categorização como página-conteúdo. Segue abaixo o exemplo do caso mais comum em chamadas para eventos científicos ao vivo (conferências e *workshops)*, em que uma página apresenta no seu início os atributos data inicial do evento e localização (país do evento) e algum termo relacionado ao conceito de evento ao vivo, como as expressões *"call for papers" (*por herança*), "conference"* ou *"workshop"*. Uma regra associada aos atributos do caso (e.g. `Slots-in-the-Beginning)` é disparada se as condições são atendidas.

```
([Date-time-in-the- beginning] of Case
      (Slots-in-the-Beginning [Initial-Date] [takes-Place-at])
      (Concepts-in-the-Beginning [live-scientific-event]))
```

- Ontologias auxiliares: Conhecimento útil de outras áreas de conhecimento. Ontologias lingüísticas, como o WordNet [Miller 95], de tempo e locais, além de outras específicas de um agente (como dados bibliográficos para o agente de artigos científicos).

4. Estudo de caso: o grupo científico

A ontologia do domínio científico [Freitas 2001] foi reusada a partir da ontologia do projeto europeu $(KA)^2$ (*Knowledge Annotation Initiative of the Knowledge Acquisition Community*) [Benjamins et al 98], refinada em vários aspectos. O principal deles foi a inclusão de classes abstratas – que não contêm instâncias –, visando abarcar classes com características comuns. Por exemplo, a classe Evento-Científico dividiu-se em duas subclasses abstratas, Evento-Científico-ao-Vivo (com subclasses Conferência e Workshop) e Evento-de-Publicação-Científica (com subclasses Jornal e Revista). Esta mudança facilitou o reconhecimento e emprestou granularidade e coerência à ontologia.

Técnicas de PLN não foram empregadas nos protótipos. Com cada agente foram realizados três testes para classificação funcional e de conteúdo de páginas e dois testes para extração de informação. Para dois dos três primeiros testes, lançou-se mão de *corpus* de páginas recuperadas de consultas a mecanismos de busca; o primeiro *corpus*, para aquisição de conhecimento (definir casos, regras e *templates*) e o segundo para teste cego. O terceiro teste foi feito acessando diretamente a Web. Dois agentes do grupo científico foram elaborados: o agente CFP, que trata páginas de chamadas de trabalhos ("*Call for papers*") de eventos científicos, como conferências e jornais, classificando-as em oito classes de páginas (as quatro citadas acima, mais Evento-

Genérico-ao-Vivo, Evento-Genérico-de-Publicação e Edição-Especial-de-Jornal e Revista) e o agente de artigos científicos, que processa páginas de artigos e documentos científicos, refinando consultas aos mecanismos de busca com palavras-chave bastante comuns em artigos: *"abstract"*, *"keywords"*, *"introduction"*, *"conclusion"*, etc. Este último classifica as páginas em artigos de *workshop*, conferência, jornal e revista, capítulo de livro e artigos genéricos, além de teses, dissertações, relatórios técnicos e de projeto.

Cada uma das oito classes citadas anteriormente é representada por um conjunto de atributos. Confore colocado na seção 3.1, para cada atributo a ser extraído, foi criada uma instância de *template*. Como exemplo, considere o seguinte trecho de uma página CFP:

> "***Paper deadline***
> *Technical Paper must be* ***submitted by****:* ***July 17, 1995***
> *(Papers must be complete for review with all references, figures etc.);*
> ***Notification of acceptance****:* ***September 4, 1995*** *(Reviewers may suggest modifications.)*"

A data "17 de julho de 1995" (em inglês, "*July 17, 1995*") refere-se à palavra-chave "*submitted by*" que expressa uma data limite para entrega de trabalhos. Isto é definido pela proximidade entre os dados e a ausência de sinais que anulem esta relação, como é o caso da mesma data (17 de julho de 1995) e a palavra-chave "*Notification of Acceptance*", onde entre os dois dados extraídos ocorre o sinal de ponto-e-vírgula que anula a relação entre eles.

4.1. Resultados

Os resultados obtidos se referem a dois conjuntos de testes. O primeiro compreende um total de quatro testes para avaliar o desempenho do MASTER-Web na classificação funcional e de conteúdo. O segundo conjunto avalia o desempenho do sistema no processo de extração de informações.

4.1.1. Classificação Funcional e de Conteúdo

A figura 2a mostra as performances dos agentes. Nela, reconhecimento indica se uma agente identificou corretamente páginas-conteúdo. A classificação de páginas-conteúdo Um quarto teste foi rodado com o agente CFP na Web, desta feita beneficiando-se da categoria funcional listas. Listas de CFPs sobre um dado assunto costumam ser mantidas por pesquisadores e organizações.

Agente CFP: Mais de 70% dos eventos eram conferências. O agente reconheceu erradamente páginas longas, geralmente sobre um assunto ou de uma comunidade (linux, XML, etc). Chamadas de eventos que não empregam o jargão comum a eventos científicos não foram reconhecidas. Listas foram detectadas pela presença de um número factível de âncoras citando palavras-chave relativas a eventos, ou por uma certa quantidade de intervalos de tempo (como 1-4 de dezembro). As listas devem ser reconhecidas com precisão, pois podem levar ao tratamento de muitas páginas inúteis.

O uso de listas melhorou a recuperação de páginas úteis entre 13 a 22% (ver Figure 2b), retornando um conjunto mais focado. Páginas de outras classes, como

mensagens e sugestões para os agentes de Organizações e de Publicações-Divisíveis (que trata Anais e Livros) foram substituídas por páginas conteúdo sob a forma de *framesets*. Nos outros testes, só um *frame* foi encontrado.

	Agente "Call for Papers"				Agente Artigos		
	Corpus Aquisição	Corpus de Teste	Teste na Web	Web usando Listas	Corpus Aquisição	Corpus de Teste	Teste na Web
Reconhecimento	97,1	93,9	96,1	96,3	93,1	82,7	87,8
Categ Funcional	93,8	93,9	93,8	95,7	96,8	94,0	95,1
Classif.Conteúdo	94,9	93,3	92,9	91,7	97,0	93,0	81,4
Págs Processadas	244	147	129	188	190	150	184

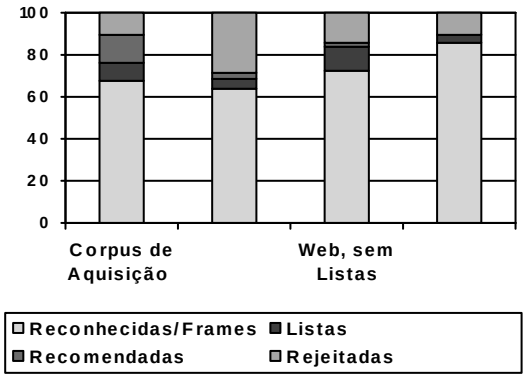


Figura 2. a) Performance dos agentes CFP e de artigos nos testes. b) Gráfico evidenciando o ganho de desempenho com o uso de listas no agente CFP.

Até o padrão das páginas rejeitadas mudou: ao invés das páginas erradas devidas à flata de precisão dos mecanismos de busca, vieram páginas gráficas iniciais de eventos. Digno de nota ainda que eventos encontrados a partir de listas não apareceram em outros testes. Estes fatos justificam a categorização funcional e o uso de listas.

Agente de artigos: Os erros no reconhecimento devem-se a artigos com poucos atributos, com atributos difíceis de identificar, - como a afiliação a uma empresa desconhecida - ou com atributos no fim do artigo. Um artigo deve possuir um conjunto mínimo destes atributos; artigos apenas com o nome do autor não foram reconhecidos, mas isto também ocorre em sistemas similares como o CiteSeer [Bollacker et al 98]. A classificação por conteúdo baseou-se em dados de publicação dos artigos, presentes no topo das páginas. Porém, mais da metade deles não traziam esses dados.

Cooperação: Efetuou-se um teste integrado, em que os agentes cooperaram. O agente CFP pediu ao agente de artigos âncoras no topo de artigos contendo conceitos como "conferência" e "jornal". Apesar de ter funcionado, apenas três páginas foram sugeridas pelo agente de artigos ao agente CFP e nenhuma página foi sugerida erradamente.

Para evidenciar que a cooperação pode ser útil, o agente CFP procurou, no atributo *Comitê de Programa*, sugestões de páginas ao futuro agente de pesquisadores. Nenhum dicionário de nomes ou técnica de extração foram empregados. O agente sugeriu 30 *links* corretos e 7 errados, um bom resultado, pois páginas de pesquisadores são menos estruturadas, portanto difíceis de ser recuperadas por mecanismos de busca.

4.1.2. Extração de Informação

Foram executados dois testes de extração de informação com o agente CFP. O primeiro foi realizado com um *corpus* de teste para aquisição de conhecimento, onde se fez ajustes (correção de instâncias e seus atributos) para maximizar a extração, enquanto que o segundo visou obter resultados para validação do desempenho do sistema sem que ajustes fossem permitidos. A figura 3 mostra uma tabela dos resultados obtidos nos testes. A cobertura é um índice quantitativo, ou seja, a porcentagem de informações

extraídas das páginas, enquanto que a precisão é um índice qualitativo, referindo-se ao número de informações que foram extraídas corretamente das páginas.

	Cobertura(%)	**Precisão(%)**
Local	68,75	68,75
Deadline	75,00	71,43
Período	78,49	57,89
Lista de Tópicos	70,00	60,00
Data de Aceitação	88,89	77,78
Total	75,60	67,06

Figura 3. Cobertura e precisão na extração de informação realizada pelo agente CFP.

As informações extraídas foram: local, *deadline* e lista de tópicos. Os resultados obtidos dependem diretamente das instâncias criadas na ontologia, ou seja, da experiência do especialista humano. A extração de informação atingiu um índice médio de cobertura de 75,6 e de precisão de 67,06, mostrando que a extração foi eficiente. É importante ressaltar que não foram usadas técnicas de aprendizado que aumentariam sensivelmente a precisão do sistema.

5. Comparação com trabalhos similares

5.1. WebKB: Classificação e Extração Baseadas em Aprendizado e Ontologias

O sistema WebKb [Craven et al 98] aprende automaticamente regras de categorização e extração integrada de páginas na Web, empregando uma ontologia do domínio com classes e relacionamentos, definida num formalismo que pode permitir inferência. As páginas da Web são representadas, com título, palavras-chave, freqüências e ponteiros.

A decisão de usar aprendizado automático depende de alguns fatores. O primeiro é uma comparação entre os custos de anotação de *corpi* e o trabalho de inspeção e aquisição do conhecimento. Existem vantagens em usar aprendizado, como velocidade e adaptabilidade, e desvantagens como legibilidade, engajamento ontológico das regras aprendidas – que tendem a ser muito específicas -, e dificuldades de aproveitar conhecimento *a priori*, de capturar regras sem introduzir porção de características para o aprendizado, e de generalizar sobre um grande número de características ou classes.

O sistema emprega uma ontologia do domínio com, apenas quatro entidades: atividades (e.g. projetos e cursos), pessoas (estudante, professor, membro do *staff*, etc), e departamentos. Relações também estão presentes, como instrutores de cursos, membros de projeto, orientadores, e outras.

Os autores do WebKB avaliam a classificação apenas através dos falsos positivos, reportando percentagens entre 73 e 83 %, exceto para as classes *membro do Staff* e *outros* (rejeitadas). Contudo, se computados os falsos negativos, a classe *outros* tem boa performance (93,6%), a classe estudante tem 43% e as outras seis classes comportam-se abaixo de 27%, baixando a média de acerto para apenas cerca de 50%. Isto leva à hipótese de que a ontologia empregada no WebKB não tenha sido abrangente o suficiente. Já a ontologia de Ciência usada pelo MASTERWeb possui classes, como projetos e produtos, que não foram usadas por dois motivos: os agentes precisam destes conceitos para suas funções, e futuros agentes que tratem delas podem ser elaborados.

Por outro lado, uma ontologia com muitas classes pode dificultar a generalização do aprendizado. Neste caso, seriam necessários mais agentes com aprendizado.

5.2. Os Sistemas CiteSeer e DEADLINER

Estes sistemas perfazem uma eficiente recuperação, filtragem e extração da Web, usando métodos estatísticos e de aprendizado combinados com conhecimento a priori.

O CiteSeer [Bollacker at al 99] é um dos mais usados na busca de artigos científicos. O sistema monitora newsgroups e editores e mecanismos de busca a partir dos termos “publications”, “papers” e “postscript”. São extraídos dados bibliográficos do artigo e da bibliografia, que atua como lista, ajudando a achar outros artigos.

O DEADLINER [Kruger et al 2000] busca anúncios de conferências, extraindo deles data inicial, final e limite, comitê, afiliação de membros do comitê, temas, nome do evento e país. A performance de reconhecimento do DEADLINER está acima de 95%, contudo, sua definição de evento é mais restritiva: todos os atributos têm de estar presentes, exceto país, além de dados de submissão. O MASTER-Web oferece mais flexibilidade e cobertura, aceitando anúncios de capítulos de livros, jornais, revistas e concursos. Os requisitos estão em casos, que são mais flexíveis.

O problema destes sistemas é que, mesmo sendo confeccionados pelo mesmo grupo de pesquisa, ambos deparam com *links* que interessam ao outro, e, não podem repassá-los. Sob o prisma de uma possível multiplicação de extratores pela Internet, isto deriva de um problema de representação de conhecimento: os dois sistemas (e outros que surgirão) não podem expressar intenções como pedir páginas ou sugerir *links*, pela falta de ontologias do domínio ou de páginas da Web. O conhecimento destes sistemas está escondido dentro de algoritmos, não sendo possível o compartilhamento deles para outros sistemas, nem a especificação de contextos em que seriam úteis. Outra vantagem está no reuso massivo da arquitetura, em que apenas parte do conhecimento tem de ser descoberto e especificado. Numa abordagem como a do CiteSeer e DEADLINER, para processar uma nova classe, um novo sistema precisa ser elaborado, sem maiores reusos.

6. Trabalhos futuros e conclusões

O projeto pode estender-se em várias direções. Novos agentes para o grupo serão desenvolvidos, como o agente de pesquisadores, e a cooperação tornar-se-á mais efetiva. Técnicas de aprendizado e PLN serão incluídas visando lidar com classes de páginas menos estruturadas. Alguma forma de checar duplicatas também será implementada.

Os agentes cognitivos, por basearem-se em modelos com conhecimento, podem comunicar-se e evitar redundância de tarefas num mesmo ambiente, Isto pode proporcionar a inauguração de uma nova era na informática distribuída, a comunicação em nível de conhecimento, dinamicamente estabelecida durante a execução, e o processamento de nichos de informação consistentes, como o domínio científico, o domínio turístico, etc. A idéia motivadora é a de que mecanismos de busca baseados em palavras-chave podem constituir a base para agentes cooperativos mais precisos e focados em domínios restritos, baseados em ontologias.

Bibliografia

L O Álvares, J S Sichman (1997). Introdução aos sistemas multiagentes. In C M B Medeiros, editor, *Jornada de Atualização em Informática (JAI'97)*, chapter 1, pages 1–38. UnB, Brasília.

R Benjamins, D Fensel and A G Pérez. (1998) Knowledge Management through Ontologies. Proc. of the 2nd International Conf. on Practical Aspects of Knowledge Management, Basel, Switzerland.

K Bollacker, S Lawrence and C L Giles. (1998) CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. Proceedings of the 2nd International ACM Conference on Autonomous Agents, USA.

M Craven, A McCallum, D DiPasquo, T Mitchell, D Freitag, K Nigam. Stephen Slattery. (1999) Learning to Extract Symbolic Knowledge from the World Wide Web. Technical Report CMU-CS-98-122. School of Computer Science. CMU, USA.

F Freitas (2001) Ontology of Science. http://protege.stanford.edu/plugins/ontologyOfScience/ontology_of_science.htm

F Freitas, G Bittencourt (2003) An Ontology-based Architecture for Cooperative Information Agents. To appear in: Proceedings of International Joint Conference of Artificial Intelligence (IJCAI`2003), Acapulco, Mexico.

A Kruger, C L Giles, F Coetze, E Glover, G Flake, S Lawrence and C Omlin. (2000) DEADLINER: Building a new niche search engine.Conf. on Information and Knowledge Management, Washington DC.

G Miller (1995) WordNet: A Lexical Database for English. Communications of the ACM. 38(11):39-41. EUA.

E. Riloff. (1994) Information Extraction as a Basis for Portable Text Classification Systems. PhD. thesis. Dept of Computer Science. Univ of Mass. at Amherst. USA.

Investigação sobre a Identificação de Assuntos em Mensagens de Chat

Stanley Loh[1,2], Daniel Lichtnow[1], Ramiro Saldaña[1], Thyago Borges[1], Tiago Primo[1], Rodrigo Branco Kickhöfel[1], Gabriel Simões[1]

[1]Universidade Católica de Pelotas (UCPEL) – Grupo de Pesquisa em Sistemas de Informação
R. Félix da Cunha, 412, Pelotas, RS – CEP 96010-000
[2]Universidade Luterana do Brasil (ULBRA) – Faculdade de Informática
R. Miguel Tostes, 101, Canoas, RS – CEP 92420-280

{lichtnow, rsaldana, thyago, rodrigok}@ucpel.tche.br, sloh@terra.com.br,
tiagoprimo@brturbo.com, gsimoes@vetorial.net

Resumo

O objetivo do presente trabalho é investigar o processo de classificaçã o de textos sobre mensagens de um chat na Web. As mensagens de chat possuem algumas particularidades como concisão, pouco cuidado com correções ortográficas e revisões, devido a serem escritas às pressas e geralmente por pessoas leigas. Neste trabalho, foram utilizados métodos simples de classificação para investigar as particularidades do processo de classificação sobre este tipo de texto.

Abstract

This paper investigates the process of classification of textual messages in a web chat. This kind of text has special characteristics, like concision, orthographic mistakes, mainly by being written in a hurry and by naïve people. In this work, simple methods were evaluated to raise some specialties of the process.

1 Introdução

Este trabalho se insere na área de classificação de textos (ou categorização). Muitos trabalhos já foram publicados nesta área, geralmente apresentando métodos novos de classificação ou técnicas para seleção de características. SEBASTIANI (2002) apresenta um *survey* sobre métodos de classificação de textos.

Entretanto, a maioria dos trabalhos realiza as avaliações dos métodos sobre textos bem escritos, como é o caso das coleções Reuters ou OHSUMED (SEBASTIANI, 2002). A primeira coleção contém textos jornalísticos, e a segunda contém títulos e resumos de textos médicos da base de dados MEDLINE. A característica principal de tais textos é que eles contêm informações bem objetivas e são escritos e revisados por profissionais (portanto, com pouquíssimos erros ortográficos). Uma dúvida que surge é se os mesmos métodos teriam o mesmo desempenho sobre textos com características diferentes.

Uns poucos trabalhos utilizaram outro tipo de coleção textual em suas avaliações, entre eles: BAKER & McCALLUM (1998), JOACHIMS (1997), LANG (1995), McCALLUM & NIGAM (1998), McCALLUM ET AL. (1998), NIGAM ET AL. (2000) e SCHAPIRE & SINGER (2000). Estes trabalhos utilizaram mensagens do *newsgroup* Usenet, uma lista de discussão na Web (disponível em http://www.cs.cmu.edu/~textlearning). Tal coleção é formada por mensagens de *e-mail*, postadas para 20 grupos de discussão diferentes. Cada grupo forma uma categoria diferente (um assunto para cada grupo). A diferença deste tipo de texto para os anteriores de Reuters e OHSUMED é que são escritos por pessoas leigas, muitas vezes com pressa e sem muito cuidado com ortografia ou erros

de digitação. A partir destes trabalhos, pode-se pensar em avaliar métodos de classificação de textos sobre mensagens de chat, que têm particularidades diferentes das mensagens de correio eletrônico, como as usadas na coleção do *newsgroup* Usenet.

Neste sentido, pode-se citar o trabalho de Khan e outros (KHAN ET AL., 2002), que analisaram mensagens de um chat mas com o objetivo de identificar relações sociais. Não foram explicados os métodos de extração de assuntos utilizados, nem foi feita nenhuma avaliação neste sentido.

O objetivo do presente trabalho é investigar o processo de classificação de textos sobre mensagens de um chat na Web. As mensagens de chat possuem algumas semelhanças com as de *newsgroups* por poderem ser escritas por qualquer pessoa, por serem escritas às pressas (a pressa é bem maior no chat que no *newsgroup*) e pelo pouquíssimo cuidado com a ortografia e erros de digitação. Mas as mensagens de chat se diferenciam principalmente por serem menores, muito mais concisas. Só para se ter uma idéia, as mensagens de *newsgroups* contêm, na maioria das vezes, mais de um período (texto entre pontos), enquanto que as de chat são geralmente formadas por apenas um período. Além disto, as mensagens de chat se caracterizam pela informalidade da linguagem utilizada.

Neste trabalho, foram utilizados métodos simples de classificação, para investigar as particularidades do processo com este tipo de texto. Não é objetivo apontar o melhor método, mas investigar como algumas características específicas de mensagens de chat podem influenciar o processo.

A identificação de assuntos em mensagens de chat tem várias aplicações, discutidas nas conclusões deste artigo. Uma delas é auxiliar sistemas de recomendação como o SisRecCol – Sistema de Recomendação para Apoio à Colaboração, o qual indica itens de uma Biblioteca Digital para participantes de um chat privado, conforme os assuntos que estão sendo discutidos (protótipo disponível em htt://gpsi.ucpel.tche.br/sisrec).

A seção 2 deste artigo apresenta os métodos utilizados para classificação das mensagens, a seção 3 discute as avaliações feitas e seus resultados, a seção 4 analisa os erros nos métodos e a seção 5 apresenta e discute as conclusões e contribuições.

2 Métodos Utilizados

O método de identificação de assuntos (*Text Mining*) funciona como um *sniffer* examinando cada uma das mensagens enviadas pelos usuários participantes do chat. Os assuntos ou temas são identificados pela comparação dos termos que aparecem nas mensagens com termos que estão relacionados a conceitos presentes numa ontologia (armazenada internamente na mesma ferramenta e descrita na próxima seção).

O método faz na verdade classificação (ou categorização) de textos, isto é, identifica assuntos nos textos presentes nas mensagens. Este método foi apresentado em LOH et al. (2000) e atingiu índices de acerto acima de 60% para textos de prontuários médicos de uma clínica psiquiátrica (o que é considerado ótimo, num domínio complexo como este). Este método está baseado em técnicas probabilísticas e, portanto, não utiliza técnicas de Processamento de Linguagem Natural para analisar a sintaxe de cada mensagem.

O algoritmo usado é baseado nos algoritmos Rocchio e Bayes (ROCCHIO, 1966; RAGAS & KOSTER, 1998; LEWIS, 1998) utilizando vetores de texto para representar textos e assuntos. O método avalia a similaridade entre o texto e um assunto usando uma

função de similaridade que calcula a distância entre os dois vetores, um representando a mensagem do chat e outro representando o assunto (da ontologia). Os vetores que representam os textos e os assuntos são compostos por uma coleção de termos onde existe um peso associado a cada termo. No caso das mensagens, o peso de cada termo é dado pela freqüência relativa de cada termo no texto, isto é, o número de ocorrências de um termo dentro de uma mensagem dividido pelo número total de termos na mesma mensagem. Já o peso de um termo em relação a um assunto representa a probabilidade que o termo tem de indicar um determinado assunto. Este peso é definido na ontologia e será melhor explicado na seção seguinte.

Na montagem dos vetores são ignoradas as chamadas *stopwords* - termos que aparecem com muita freqüência e que não tem relevância na identificação dos assuntos de uma discussão, tais como preposições e artigos, por exemplo. Feita a montagem dos vetores, os dois são comparados por meio de uma função *fuzzy* de similaridade. O método utilizado multiplica os pesos dos termos que estão presentes nos dois vetores, sendo que a soma destes produtos, limitada a 1, é o grau de similaridade existente entre a mensagem e o assunto. Este grau determina qual a probabilidade do assunto estar presente na mensagem. Um limiar mínimo (*threshold*) é usado para cortar graus indesejados, abaixo do qual é improvável que o assunto esteja presente na mensagem. Este limiar é estabelecido por especialistas humanos na configuração do sistema e funciona para todas as sessões. O limiar ainda está sendo testado.

O método é baseado no índice de relevância, proposto por RILOFF & LEHNERT (1994). O índice de relevância é "um conjunto de características que juntas podem predizer com confiança a descrição ou existência de um evento". Partindo desta premissa, o método considera que alguns termos presentes nas mensagens do chat podem portanto indicar a presença de um assunto com um grau de certeza. Conseqüentemente o processo de raciocínio *fuzzy* deverá avaliar a probabilidade de um assunto estar presente em um texto, analisando a intensidade desta s indicações (presença de termos). Isto significa que, se as palavras que descrevem um assunto *c* aparecem em um texto, existe uma probabilidade alta do assunto *c* estar presente no texto. A soma das indicações deve resolver problemas de ambigüidade. Por exemplo, a presença do termo '*inteligência*' gera uma ambigüidade por se referir tanto ao tema "inteligência artificial" quanto ao tema "inteligência competitiva", mas a presença de outros termos ajuda a resolver tal conflito.

A abordagem pode ser considerada sob o paradigma de processamento estatístico de linguagem natural, conforme classificação de KNIGHT (1999), uma vez que analisa freqüência de palavras e probabilidades.

Inicialmente estão sendo testados 2 métodos derivados, a saber:

a) um método que analisa todas as mensagens enviadas e, para cada uma delas, aceita como verdadeiro apenas o assunto que tem maior grau; neste método, se dois assuntos são identificados com o mesmo grau e se existe entre eles uma hierarquia (pai e filho), então o assunto mais específico é utilizado; caso o empate ocorra entre assuntos que estão em um mesmo nível da hierarquia, então um dos assuntos será tomado como verdadeiro de forma aleatória; uma variação seria permitir que vários assuntos fossem identificados para cada mensagem (formando um *ranking* pelo grau ou probabilidade);
b) um método que avalia um conjunto de mensagens para identificar o assunto sendo tratado na discussão; no método anterior, cada mensagem era analisada individualmente

para se identificar o assunto; neste, os pesos ou graus associados a cada assunto para cada mensagem vão sendo somados sendo um assunto identificado somente quando a soma dos pesos passa um determinado limiar (configurado manualmente) .

Conceito Identificado	Peso	Mensagem enviada
PROJETO DE BANCO DE DADOS	0.000900	éé, rodrigo como funcia a esquema, a tabela q grava os pesos fica sempre gravando? ou ela e limpa em determinado ponto?
REDES DE COMPUTADORES	0.002365	tô no notebook, mas pode ser minha conexão cable modem
REDES DE COMPUTADORES	0.001152	Abri as recomendações de redes parece que está bem
PROJETO DE BANCO DE DADOS	0.000900	uma chave estrangeira pode ser ao mesmo tempo uma chave primária?

Figura 1: Trecho extraído de uma sessão no chat

Para exemplificar a diferença entre estes 2 métodos, note-se a figura 1, contendo um trecho extraído de uma sessão no chat. Pelo primeiro método, tem-se na primeira coluna o assunto identificado para cada mensagem enviada (e o peso associando o assunto à mensagem). Neste caso, cada mensagem individualmente gera um assunto identificado. Pelo segundo método, os pesos gerados para cada mensagem devem ser somados, agrupados por assunto. Assumindo que o limiar mínimo fosse de 0,001 (que é o limiar que está sendo testado no momento), a primeira mensagem não geraria nenhum assunto. Já a segunda e a terceira mensagens gerariam assuntos de forma individual. A quarta mensagem geraria o assunto "Projeto de Banco de Dados" pela soma (o peso do assunto identificado nela com o peso da primeira mensagem, que também identificou este mesmo assunto).

Os métodos foram implementados utilizando as tecnologias livres como linguagens de programação PHP e Javascript, banco de dados MySql, servidor *Web* Apache e sistema operacional Linux. Estes métodos compõem um protótipo de sistema de recomendação, que encontra-se disponível no endereço http://gpsi.ucpel.tche.br/sisrec.

2.1 A Ontologia

O sistema utiliza uma ontologia de domínio para classificar documentos, para identificar temas nas mensagens e para traçar o perfil dos usuários. Uma ontologia de domínio (*domain ontology*) é uma descrição de "coisas" que existem ou podem existir em um domínio (SOWA, 2002) e descreve o vocabulário relacionado ao domínio em questão (GUARINO, 1998).

Neste trabalho, a ontologia foi implementada como um conjunto de assuntos em uma estrutura hierárquica (um nó raiz, e nós pais e filhos), onde cada assunto tem associado a si uma lista de termos e seus respectivos pesos, que ajudam e identificar o assunto presente nos textos das mensagens. Os pesos associados aos termos determinam a importância relativa ou a probabilidade de um determinado termo identificar o assunto em um texto.

Para gerenciar e manter a ontologia foram implementadas algumas ferramentas que permitem sua visualização (hierarquia de assuntos, termos associados aos assuntos), a inclusão e a remoção de assuntos e de termos e a modificação dos pesos dos termos associados aos assuntos. Na implementação atual, a ontologia está voltada para a área de Ciência da Computação, sendo os assuntos baseados na classificação da ACM (www.acm.org), mas novas sub-áreas foram acrescentadas. Entretanto, outras ontologias podem ser adicionadas.

A ontologia foi criada de forma semi-automática. A seleção de assuntos (áreas e sub-áreas da hierarquia) foi feita manualmente por especialistas na área de Computação. Após, ferramentas automatizadas foram utilizadas para identificar termos que pudessem indicar cada assunto, seguindo um processo tipicamente de aprendizado de máquina (*machine learning*). Neste processo, especialistas nos assuntos selecionaram documentos eletrônicos de cada assunto (aproximadamente 100 documentos para cada assunto) e uma ferramenta de software identificou os termos mais relevantes e determinou o peso de cada termo. Este peso foi calculado com base na freqüência do termo dentro dos documentos e também avaliando o número de documentos daquele assunto onde o termo aparecia. Este é um procedimento típico do método Rocchio, sendo que o vetor gerado chama-se centróide.

Uma revisão dos termos e pesos foi feita por especialistas nas áreas relacionadas, observando os termos que apareciam em mais de um assunto (considerados termos genéricos, os quais tiveram seu peso diminuído) e procurando normalizar os pesos (os maiores pesos em cada assunto deveriam estar em patamares semelhantes).

A ontologia possui termos em português e inglês. Para tanto, foi necessário gerar termos nas duas línguas, quando o processo automático não conseguiu identificá-los. Como não se usa nenhum tratamento de radicais (*stemming*), foi necessário gerar manualmente as variações lingüísticas (número, gênero e as principais conjugações verbais).

3 Avaliação Formal do Método

O método de identificação de assuntos em mensagens de chat foi avaliado de duas formas. A primeira foi uma avaliação feita de forma *offline*, tomando como entrada resumos (abstracts) e parágrafos de textos (artigos científicos) selecionados manualmente por assunto. A avaliação comparou os textos (resumos e parágrafos) de forma completa em relação a frases (períodos) extraídos destes textos. O objetivo desta avaliação era saber o grau de acerto do método quando textos bem escritos e bem objetivos eram utilizados como entrada para representar mensagens de um chat. Neste caso, procurou-se observar se o texto todo ou parte dele (frases) poderiam gerar resultados diferentes.

A segunda avaliação foi feita *online* sobre as mensagens enviadas pelo chat, durante sessões reais de discussão de grupos de pesquisa ou comunidades virtuais.

3.1 Avaliação Offline

Foram selecionados 15 artigos científicos de diversas área de Computação, coletados a partir da Biblioteca Digital Citeseer ou ResearchIndex (www.researchindex.com). Destes artigos, foram extraídos os resumos (abstracts) e 2 parágrafos do meio de cada artigo (escolha aleatória). Procurou-se observar o nível de acerto do método nas seguintes situações:

a) quando o resumo todo era submetido como entrada (admitindo uma mensagem extensa no chat);
b) quando cada frase do resumo era submetida individualmente como entrada (uma por vez, gerando cada frase uma avaliação diferente, ou seja, um assunto identificado para cada uma);
c) quando cada parágrafo extraído do artigo era submetido como entrada;
d) quando cada frase dos 2 parágrafos de cada texto era submetida como entrada.

Cada artigo foi previamente associado por especialistas a um assunto da ontologia, e para fins de avaliação dos métodos, este assunto foi assumido como o correto tanto para os resumos e parágrafos extraídos deste texto, quanto para as frases extraídas.

A saber, foram avaliadas no total 78 frases extraídas de resumos e 418 frases extraídas de parágrafos. Os resultados são apresentados na tabela 1. Os textos maiores geraram melhores resultados. Isto já era esperado, uma vez que o método é probabilístico e, portanto, identifica melhor o assunto quando existe um maior número de características presentes. Acredita-se que os resumos geraram melhores resultados que os parágrafos por terem informações mais abrangentes sobre a área, enquanto que os parágrafos poderiam tratar mais de detalhes do artigo.

Tabela 1: Resultados da Avaliação Offline

Tipo de entrada	% de acertos
Resumos	91,66%
Frases dos resumos	60,97%
Parágrafos	83,33%
Frases dos parágrafos	58,73%

3.2 Avaliação Online e Comparação entre Métodos

Para a avaliação *online*, foram selecionadas 3 sessões de discussão, onde grupos de pesquisa utilizaram o chat. As mensagens enviadas para o chat foram analisadas pelo sistema de identificação de assuntos conforme os métodos descritos anteriormente. As mensagens da primeira e da segunda sessão foram analisadas com o primeiro método, que identifica assunto para cada mensagem individualmente. Já as mensagens da terceira sessão foram analisadas com o segundo método, que procura identificar o assunto pela soma dos pesos das mensagens (agrupadas por assunto).

Os próprios participantes das sessões analisaram os assuntos identificados para as mensagens e decidiram o que estava correto ou errado, bem como as mensagens que deveriam ter gerado assunto e não o fizeram.

A primeira sessão teve um total de 168 mensagens enviadas para o chat, sendo que em 48 mensagens foi identificado um assunto, mas em 120 mensagens não foi possível identificar nenhum assunto. Destas 120 mensagens, 9 mensagens deveriam ter permitido identificar algum assunto. Das 48 mensagens, 18 permitiram identificar o assunto correto.

A segunda sessão teve um total de 184 mensagens enviadas para o chat, sendo que em 52 mensagens foi possível identificar um assunto, mas em 132 mensagens não foi identificado nenhum assunto. Em 26 mensagens, foi identificado o assunto correto. Novamente em 9 mensagens deveria ter sido identificado um assunto e isto não ocorreu.

Na primeira sessão, houve uma precisão de 37,5% (proporção de assuntos corretamente identificados) e abrangência de 66,6% (proporção de mensagens com assunto corretamente identificado em relação ao total de mensagens que deveriam gerar assunto). Já na segunda sessão, a precisão ficou em 50%, sendo a abrangência igual a 74,3%. Uma das explicações para a melhora de precisão é que a segunda sessão foi mais técnica, isto é, poucas mensagens estavam relacionadas a aspectos administrativos do grupo.

A avaliação do segundo método (terceira sessão) levou em conta um conjunto de mensagens para identificar um assunto, e não mensagens individualmente. Foi utilizado um limiar para somar os pesos, como explicado anteriormente. Procurou-se determinar, além do grau de acerto em geral, qual limiar gerava melhores resultados. A terceira sessão teve

um total de 374 mensagens, sendo que em 258 não foi possível identificar nenhum conceito. Em 116 mensagens, foi possível identificar um assunto. Entretanto, algumas destas mensagens tinham associado a elas um peso abaixo do limiar. Portanto, para fins de avaliação deste segundo método, somente foram consideradas as mensagens que geraram a identificação de um assunto, isto é, quando a soma dos pesos ficava acima do limiar estabelecido.

Com um limiar igual a 0,001, foram identificados assuntos para 83 mensagens, isto é, 83 mensagens geraram um assunto quando a soma ultrapassou o limiar. Portanto, das 116 mensagens, 33 não ultrapassaram o limiar pela soma. Das 83, 56 tiveram o assunto correto identificado. Em 9 mensagens deveria ter sido identificado um assunto mas nada foi identificado. Assim, a precisão ficou em 67,5% e a abrangência em 86,1%. Com o limiar em 0,005, foram identificados assuntos em 63 mensagens, com 54 corretas, e tendo 9 mensagens sido deixadas de fora. Neste caso, a precisão melhorou para 85,7% e a abrangência baixou um pouco para 85,7%. Aumentando o limiar para 0,01, 54 mensagens geraram assuntos (45 corretas e 9 deixadas de fora), resultando numa precisão de 83,3%, mas a abrangência caiu levemente para 83,3%.

Deste resultado, conclui-se que o limiar 0,005 poderia ser utilizado neste método que considera a soma dos pesos. As mensagens para as quais deveria ter sido identificado um assunto e não foi, não foram deixadas de fora por causa do limiar, mas sim por erros de outro tipo.

Comparando os métodos, o que se nota é que o método que utiliza a soma dos pesos para identificar um assunto (usado na terceira sessão) aumenta bastante a precisão, pois somente identifica um assunto quando várias mensagens com pesos médios forem enviadas ou quando uma mensagem com peso alto é enviada (simulando uma análise de contexto).

4 Análise dos Erros

A partir dos experimentos de avaliação dos métodos utilizados, procurou-se analisar causas dos possíveis erros na identificação de assunto. Uma conclusão a que se chegou é que erros ortográficos, gírias e abreviaturas tendem a confundir os métodos de identificação dos assuntos. Um corretor ortográfico e a inclusão na ontologia dos demais termos que puderem ser previstos (os mais comuns) devem minimizar tais problemas.

Pôde-se notar também a presença de muitas expressões fora do contexto técnico da discussão (como *"oi"*, *"o fulano entrou no chat"*) ou que retrucavam ("*que acham*?", *"por quê*?"). Estas, em sua maioria, serviram para identificar nenhum assunto. Entretanto, algumas identificaram assuntos errados, devido a esta economia de palavras.

Uma das principais características das mensagens de um chat é serem concisas, para se ganhar tempo. A conseqüência é que há muita informação subentendida. Por exemplo, quando o grupo estava discutindo sobre "desempenho no chat devido a problemas de conexão" (assunto "redes de computadores"), muitas mensagens utilizavam somente o termo "redes", admitindo que o grupo já entendia seu significado (sem confundir com "redes neurais" ou outro assunto). A conclusão é que métodos que analisem o contexto devem ser utilizados. Neste trabalho, foi comprovado que um método que faz tal análise (mesmo apesar de não ser uma análise tão complexa) tende a melhorar a precisão (lembrando que o segundo método, que fazia a análise de um conjunto de mensagens, foi

melhor que o primeiro). Entretanto, concluiu-se que o esquema de soma dos pesos não é o melhor, uma vez que, quando a soma atinge um limiar, todas as mensagens posteriores sobre o mesmo tema irão gerar o assunto. Uma opção a ser testada futuramente é utilizar grupos de mensagens por janelas ou por tempo, para avaliar a soma (a soma só seria avaliada sobre as últimas N mensagens).

Outra constatação é que mensagens com peso baixo podem indicar a falta de um assunto específico na ontologia. Este foi o caso do uso do termo "recomendações", que indica o assunto "sistemas de informação" com peso baixo. Entretanto, poder-se-ia criar um assunto "filho" (especialização), onde este termo teria um peso maior.

A partir da análise dos conceitos erroneamente identificados, também foi possível notar que havia falhas nos pesos dos termos associados a alguns conceitos na ontologia. Termos de significado muito genérico tendem a induzir a erros se tiverem pesos muito altos (por exemplo: projeto, sistema, técnicas). Os pesos destes termos foram diminuídos manualmente na ontologia. Uma implementação futura deverá analisar os pesos dos termos na ontologia, de forma automática, para diminuir o peso de termos que aparecem em muitos assuntos, conforme sugestão de SALTON & McGILL (1983). Outra maneira de minimizar tal problema é cuidando para que os termos de maior peso em cada conceito estejam na mesma faixa de valor (normalizados).

Outro erro comum foi identificar um assunto "filho" (mais específico) quando deveria ser identificado o conceito "pai" (exemplo: o termo "tabelas" levou ao conceito "projeto de banco de dados" ao invés de "banco de dados"). A solução encontrada foi diminuir nos conceitos "filhos" o peso de termos que são mais genéricos, ou seja, identificam melhor o conceito "pai". Está sendo planejada uma ferramenta que encontra palavras que aparecem em conceitos "pai" e "filho", para serem apresentadas a um especialista que, manualmente, poderá modificar os seus pesos.

Um caso especial encontrado a partir da análise dos erros é o de mensagens com vários assuntos (ex: uma mensagem citava "inteligência artificial", "arquitetura de computadores" e outras sub-áreas). Um termo com peso alto num destes assuntos influencia o assunto final identificado. Uma solução pode ser equiparar os pesos nos assuntos (normalização). Mas a questão principal é que somente um assunto vai ser identificado (resposta final do método), quando na verdade o correto seria indicar os vários assuntos presentes (lembrando que os métodos podem detectar vários assuntos, mas somente identifica como assunto correto o de maior peso). Uma curiosidade é que mensagens com vários assuntos (corretamente identificados) tinham pesos semelhantes para os assuntos detectados. Isto poderia ser utilizado pelos métodos para identificar vários assuntos numa mesma mensagem (quando diversos assuntos forem detectados com peso acima do limiar então é porque realmente existem vários conceitos sendo discutidos na mensagem ou num grupo de mensagens).

Por fim, o sistema descrito neste artigo separa numa lista as palavras que apareceram no chat mas que não estavam presentes na ontologia nem eram "*stopwords*". Analisando-as, notou-se haver expressões que realmente não deveriam estar na ontologia (como "xi", "haha", nomes próprios, gírias, erros ortográficos e números), mas também puderam ser identificadas novas *stopwords* (verbos genéricos, por exemplo) e termos bastante significativos (exemplo "distribuídos"), que ficaram de fora da ontologia. Fica claro que a ontologia deve ser revisada e uma das formas é coletar os termos usados no chat

e que não se encontram na ontologia.

5 Conclusões

Este trabalho avaliou a identificação de assuntos em mensagens de chat. Para tanto, foram implementados, avaliados e comparados dois métodos probabilístico s.

A principal contribuição do artigo (e também sua diferença para outros trabalhos já comentados) está em avaliar o processo de classificação de textos curtos, concisos e escritos rapidamente, como é o caso das mensagens de chat. Só para constar, as mensagens do newsgroup Usenet, utilizadas em alguns trabalhos, tem em média 124,7 palavras, já excluindo as chamadas *stopwords*. Nas mensagens de chat analisadas, encontrou-se uma média de 3,3 palavras por mensagem, sem *stopwords*, e uma média de 6,35 palavras contando as *stopwords*. A concisão das mensagens de chat é um empecilho aos métodos de classificação de texto. Entretanto, foi demonstrado que métodos simples (baseados em estatística e que não utilizam análise sintática) podem consegui r um bom nível de precisão e abrangência (85,7% e 87,5%, respectivamente, no melhor caso).

Este bom desempenho ainda poderia ser melhor se erros ortográficos e de digitação, bem como gírias e abreviaturas pudessem ser identificados e corrigidos ou transformados para termos correspondentes na ontologia. A análise dos erros gerados permitiu identificar algumas características das mensagens de chat que podem influenciar o processo.

Outro fator que influencia o desempenho dos métodos de classificação é a qualidade da ontologia. Na ontologia utilizada neste trabalho, havia uma média 145 palavras em cada vetor que definia um assunto, gerando um vocabulário com 2874 palavras (mais 443 termos considerados *stopwords*), em português e inglês.

Se comparada aos vocabulários utilizados em outros trabalhos, este quantia é pequena. Por exemplo, BAKER & McCALLUM (1998) utilizaram um vocabulário de 62258 palavras, sem *stopwords*, sem tratamento de *stemming* e excluindo palavras que apareciam uma vez só. Já McCALLUM,& NIGAM (1998) utilizaram um total de 22958 palavras, sem *stemming* e sem as que apareciam só uma vez. McCALLUM ET AL. (1998) usaram um vocabulário de 52309 palavras, sem *stemming*, com lista de *stopwords*, mas removendo palavras que apareciam só uma vez. Entretanto, os trabalhos citados usaram parte das próprias mensagens como treino. A ontologia utilizada nos experimentos apresentados no presente artigo foi construída a partir de artigos científicos e não de mensagens. Somente parte do vocabulário da ontologia se fez presente nas mensagens analisadas. Para se ter uma idéia, a primeira sessão teve 534 palavras diferentes, sem contar *stopwords*. Já na terceira sessão, apareceram 374 palavras diferente s, sem considerar *stopwords*. Por curiosidade, vale salientar que as palavras mais freqüentes apareceram no máximo 11 vezes numa sessão.

A aplicação dos resultados deste trabalho poderá melhorar o processo de identificação de temas em mensagens de chat. Isto tem conseqüências diretas sobre sistemas que analisam discussões em chats, tais como:

- sistemas de identificação de especialistas ou análise de *expertise*: para encontrar pessoas autoridades em determinado assunto ou simplesmente identificar quem conhece algo sobre algum assunto;
- sistemas de recomendação: para indicar itens de forma sensível ao contexto

(ofertas personalizadas conforme interesse de cada pessoa que participa do chat);

- sistemas de publicidade online (*adverstising*): para apresentar informações personalizadas de acordo com assuntos sendo discutidos no chat.

6 Agradecimentos

O presente trabalho foi realizado com o apoio do CNPq, uma entid ade do Governo Brasileiro voltada ao desenvolvimento científico e tecnológico.

7 Referências Bibliográficas

BAKER, L. D. & McCALLUM, A. K. 1998. Distributional clustering of words for text classification. IN: Proceedings ACM International Conference on Re search and Development in Information Retrieval, SIGIR-98, 21., Melbourne, 1998. p.96-103.

GUARINO, Nicola. 1998. Formal Ontology and Information Systems. In: International Conference on Formal Ontologies in Information Systems - FOIS'98, Trento, Itália, Junho de 1998. p. 3-15

JOACHIMS, T. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. IN: Proceedings International Conference on Machine Learning, ICML-97, Nashville, 1997. p.143-151.

KHAN, Faisal M. ET AL. 2002. Mining chat-room conversations for social and semantic interactions. Technical Report, LU-CSE-02-011, Lehigh University.

KNIGHT, Kevin. 1999. Mining online text. Communications of the ACM, v.42, n.11, p.58-61.

LANG, K. 1995. NewsWeeder: learning to filter netnews. IN: Proceedings International Conference on Machine Learning, ICML-95, 12., Lake Tahoe, 1995. p.331-339.

LEWIS, David D. 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In: European Conference on Machine Learning, Chemnitz, Alemanha, 1998. p.4-15. (Lecture Notes in Computer Science, v.1398).

LOH, S.; WIVES, L. K.; OLIVEIRA, J. P. M. 2000. Concept-based knowledge discovery in texts extracted from the Web. ACM SIGKDD Explorations, v.2, n.1, Julho de 2000, p. 29-39.

McCALLUM, A. K. & NIGAM, K. 1998. Employing EM in pool-based active learning for text classification. IN: Proceedings International Conference on Machine Learning, ICML-98, Madison, 1998. p.350-358.

McCALLUM, A. K.; ROSENFELD, R.; MITCHELL, T. M.; NG, A. Y. 1998. Improving text classification by shrinkage in a hierarchy of classes. IN: Proceedings International Conference on Machine Learning, ICML-98, Madison, 1998. p.359-367.

NIGAM, K.; McCALLUM, A. K.; THRUN, S.; MITCHELL, T. M. 2000. Text classification from labeled and unlabeled documens using EM. Machine Learning, v. 39, n.2/3, p.103-134.

RAGAS, Hein & KOSTER, Cornelis H. A. 1998. Four text classification algorithms compared on a Dutch corpus. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval, Melbourne, 1998, p.369-370.

RILOFF, Ellen & LEHNERT, Wendy. 1994. Information extraction as a basis for high-precision text classification. ACM Transactions on Information Systems, v.12, n.3, Julho de 1994, p.296-333.

ROCCHIO, J. J. 1966. Document retrieval systems - optimization and evaluation. Tese (Doutorado)- Harvard University, Cambridge.

SALTON, G. & McGILL, M. J. 1983. Introduction to modern information retrieval. New York: McGraw-Hill, 1983.

SEBASTIANI, Fabrizio. 2002. Machine learning in automated text categorization. ACM Computing Surveys, v.34, n.1, Março de 2002.

SCHAPIRE, R. E. & SINGER, Y. 2000. BoosTexter: a boosting-based system for text categorization. Machine Learning, v. 39, n.2/3, p.135-168.

SOWA, John F. 2002. Building, sharing, and merging ontologies. Disponível em http://www.jfsowa.com/ontology

Geração de Impressão Digital para Recuperação de Documentos Similares na *Web*

Álvaro R. Pereira Jr[1], **Nivio Ziviani**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627 – 31270-010
Belo Horizonte – Minas Gerais

{alvaro, nivio}@dcc.ufmg.br

Abstract. *This paper presents a mechanism for the generation of the "fingerprint" of a Web document. This mechanism is part of a system for detecting and retrieving documents from the Web with a similarity relation to a suspicious document. The process is composed of three stages: a) generation of a fingerprint of the suspicious document, b) gathering candidate documents from the Web and c) comparison of each candidate document and the suspicious document. In the first stage, the fingerprint of the suspicious document is used as its identification. The fingerprint is composed of representative sentences of the document. In the second stage, the sentences composing the fingerprint are used as queries submitted to a search engine. The documents identified by the URLs returned from the search engine are collected to form a set of similarity candidate documents. In the third stage, the candidate documents are "in-place" compared to the suspicious document. The focus of this work is on the generation of the fingerprint of the suspicious document. Experiments were performed using a collection of plagiarized documents constructed specially for this work. For the best fingerprint evaluated, on average* 87.06% *of the source documents used in the composition of the plagiarized document were retrieved from the Web.*

Resumo. *Este artigo apresenta um mecanismo para geração da "impressão digital" de um documento da Web. Esse mecanismo é parte de um sistema para detectar e recuperar documentos que tenham sido plagiados da Web, sendo similares a um dado documento suspeito. O processo é composto de três etapas: a) geração de uma impressão digital do documento suspeito, b) coleta de documentos candidatos da Web e c) comparação entre cada documento candidato e o documento suspeito. Na primeira etapa, a impressão digital do documento suspeito é usada para identificá-lo. A impressão digital é constituída por um conjunto de frases mais representativas do documento. Na segunda etapa, as frases que constituem a impressão digital são usadas como consultas e submetidas para uma máquina de busca. Os documentos identificados pelas URLs da resposta da pesquisa são coletados e formam um conjunto de documentos candidatos à similaridade. Na terceira etapa, os documentos candidatos são localmente comparados com o documento suspeito. O foco deste trabalho está na geração da impressão digital do documento plagiado. Experimentos foram realizados sobre uma coleção de documentos plagiados construída especialmente para este trabalho. Para a impressão digital de melhor resultado, em média* $87,06\%$ *dos documentos usados na composição do documento plagiado foram recuperados da Web.*

1. Introdução

Com a Internet a sociedade tem praticado plágio com mais facilidade. Desde escolas de primeiro grau até cursos de pós-graduação, a facilidade de se efetuar um *download* e copiar a informação encontrada tem levado a uma epidemia de plágio digital. Talvez o problema mais alarmante desta epidemia de plágio digital seja a contribuição para que o plágio cada vez mais faça parte de nossa cultura educacional. Estudantes que estão crescendo com a Internet muitas vezes não estão percebendo que estão praticando o plágio. Passa a ser muito natural a ação de se "copiar" e "colar". Os estudantes estão se acostumando a somente repetir o que alguém já fez, sem criatividade, inovação e, principalmente, sem aprendizado, pois não foi ele quem fez.

Recuperar documentos que possuam o conteúdo desejado por um usuário é uma tarefa complexa, principalmente em grandes repositórios de documentos como a *Web*. Esta tarefa é função das *máquinas de busca*, que mantêm páginas da *Web* em sua base de documentos. Toda a base de documentos da máquina de busca fica indexada em forma de uma estrutura de dados chamada arquivo invertido, que permite a realização de consultas. O usuário entra com palavras chaves relacionadas à resposta que gostaria de obter e através de uma medida de similaridade entre os termos da consulta e cada documento indexado, os documentos de maior similaridade são retornados. Para o presente trabalho, o problema continua a ser a recuperação de documentos em um grande repositório de documentos. No entanto, a consulta não é mais por palavras chaves, mas sim por um documento inteiro.

Este trabalho apresenta um mecanismo capaz de detectar e recuperar documentos da *Web* que possuam uma relação de similaridade com um dado documento suspeito, ou seja, tenham sido plagiados da *Web*. O processo é realizado em três etapas principais. A primeira etapa compreende a retirada da impressão digital do documento. A impressão digital representa e identifica o documento suspeito. É composta de frases do texto, que são utilizadas na segunda etapa do processo. A segunda etapa tem o objetivo de coletar da *Web* documentos candidatos a apresentarem uma relação de similaridade com o documento suspeito. Cada frase da impressão digital é utilizada como consulta em um sistema de busca que retorna os documentos que compõem a base de documentos candidatos à similaridade. Na terceira etapa, cada documento candidato é comparado com o documento suspeito. O foco deste trabalho está na etapa de geração da impressão digital, que será detalhada na seção 2. As demais etapas serão apresentadas de forma sucinta, na seção 3.

A avaliação do processo desenvolvido se deu pela capacidade do sistema em recuperar os documentos usados para compor o documento suspeito. Para que a avaliação pudesse ser realizada desenvolvemos um sistema gerador de documentos plagiados, capaz de compor um documento utilizando trechos de diferentes documentos coletados da *Web*. O sistema retorna as URLs[1] dos documentos usados na composição do documento plagiado. Verificamos que, para a melhor impressão digital avaliada, em $61,53\%$ dos casos, todos os documentos da composição foram recuperados e que somente em $5,44\%$ dos casos o desempenho foi menor que 50%. Para esta impressão, em média $87,06\%$ dos documentos foram recuperados da *Web*.

Desde 1994 vários mecanismos de verificação de similaridade entre documentos foram propostos, usando diferentes modelos e com diferentes finalidades. A ferramenta SIF [Manber, 1994] foi a pioneira, e tratava o problema da similaridade não somente para documentos, mas arquivos binários em geral. A ferramenta COPS (*COpy Protection System*) [Brin et al., 1995] e as diferentes versões do SCAM (*Stanford Copy Analysis Mechanism*) [Shivakumar and Garcia-Molina, 1995, Garcia-Molina et al., 1996,

[1] URL (*Uniform Resource Locator*) é o identificador único de um documento na *Web*, o seu endereço.

Garcia-Molina et al., 1998] são resultados de um dos maiores estudos realizados sobre detecção de cópias em grandes repositórios de documentos. A primeira versão do SCAM abordou o problema considerando o repositório de documentos localmente. As últimas versões funcionavam considerando a *Web* como sendo o repositório de documentos. [Pereira-Jr, 2004] apresenta e discute estes e alguns outros mecanismos de detecção de cópias já propostos.

2. Geração da Impressão Digital

A primeira etapa do sistema consiste em gerar uma impressão digital para o documento a ser pesquisado. O problema está em definir as características dessa impressão digital, uma vez que cada frase da impressão é posteriormente usada como uma consulta na etapa de pesquisa e coleta.

Ao buscar definir as impressões digitais, devemos lembrar que o objetivo não é procurar na *Web* pelos exatos documentos que tiveram a impressão digital obtida. O objetivo neste trabalho é usar a impressão digital para buscar por vários documentos que possam ter sido usados na composição do documento suspeito. Desta forma, realizar pesquisas por uma lista de termos espalhados pelo texto, ou pelos termos de maior frequência no documento, poderia resultar em um baixo desempenho. Isto ocorreria porque as listas de termos mais frequentes dos documentos usados na composição do documento suspeito certamente não seriam as mesmas.

Seis diferentes impressões digitais foram estudadas e implementadas. A maioria das impressões digitais utilizadas são compostas por uma lista sequencial de termos do texto, que chamaremos de *frases*, mesmo que muitas vezes estas listas não tenham um sentido semântico. Em alguns casos foram usados termos específicos como âncoras no texto, e cada frase foi formada tomando o mesmo número de termos à esquerda (incluindo o próprio termo) e à direita do termo âncora.

Para cada uma das impressões digitais propostas temos opções de variar a granularidade e a resolução da mesma. A *granularidade* é medida de acordo com a quantidade de termos contidos em cada frase da impressão digital. Todas as frases de uma mesma impressão digital têm a mesma granularidade. A *resolução* é medida pela quantidade de frases a serem obtidas para compor a impressão digital.

Uma vez que cada frase da impressão digital será uma consulta no sistema de busca, a maior granularidade considerada foi de dez termos, número máximo aceito pela maioria das máquinas de busca. Pelo mesmo motivo, a resolução deve ser a menor possível, implicando em menos requisições à máquina de busca e menos páginas coletadas para compor a base de documentos candidatos. Os métodos estudados são apresentados a seguir:

1. **Termos mais frequentes – TF**
   Uma impressão digital contendo os termos que mais ocorrem no documento. Sua resolução é sempre de uma frase, podendo variar a granularidade.
2. **Frases com termo incorreto – FTI**
   A implementação desta impressão digital foi motivada pela intuição de que frases que envolvam termos com erros ortográficos representam bem o documento, uma vez que acredita-se ter maior probabilidade de não existirem outros documentos com os mesmos termos incorretos. Utilizando o programa "ispell" da GNU[2], todos os termos que não fazem parte do dicionário da língua portuguesa são gerados

[2] GNU é um projeto de gerenciamento de um ambiente para desenvolvimento de software livre – http://www.gnu.org

e ordenados do termo de maior comprimento para o de menor comprimento. Assim, é dada menor prioridade para termos curtos, que podem ser apenas siglas não encontradas no dicionário utilizado. Os termos no topo da lista funcionam como âncoras no texto, para retirada das frases que irão compor a impressão digital.

3. **Frases espalhadas constantes – FEC**
   Frases espalhadas no texto, equidistantes umas das outras, são usadas para formar a impressão digital do documento. Independente do tamanho do texto, sempre o mesmo número de frases são obtidas, mantendo a resolução constante.
4. **Frases espalhadas proporcionais – FEP**
   Como a impressão FEC, porém a resolução é proporcional à quantidade de caracteres do documento, calculada de acordo com a equação:
   $res = k \times \log(qtdCarac/10)$, onde $qtdCarac$ é a quantidade de caracteres do documento, k é uma constante e res a resolução.
5. **Frases com termos mais frequentes – FTF**
   É gerada uma lista com os termos mais frequentes, que são usados como âncoras no texto para retirada de frases.
6. **Frases com termos menos frequentes – FTMF**
   A lista utilizada é a de termos menos frequentes, que também são usados como âncoras na retirada de frases. Como na maioria dos casos existem muitos termos com frequência um, os termos de maior comprimento são escolhidos.

2.1. Exemplo de Impressão Digital

Como exemplo, vamos considerar o trecho de texto[3], da figura 1 mantido com erros ortográficos, gramaticais e frases mal elaboradas, como sendo o documento da consulta no qual queremos retirar as diferentes impressões digitais. Vamos considerar ainda a granularidade sendo de quatro termos e a resolução de duas frases.

> O movimento insurrecional de 1789 em Minas Gerais teve característica marcantes que o fizeram distinguir-se das outras tentativas de independência, ele foi mais bem elaborado preparado que a Inconfidência Baiana de 1798 e a Pernambucana de 1801. Os Mineiros que lideraram a conspiração de 1785-1789 tinham bem em vista a Independência Global do Brasil, e não uma republica em Minas Gerais. O plano mineiro era em iniciar a revolta por Minas Gerais, e estendê-la ao Rio de Janeiro e em seguida as demais Capitanias, o produto não foi produto da mente de ninguém em particular, nasceu das condições estruturais da sociedade brasileira.

Figura 1: Exemplo de texto plagiado

A tabela 1 mostra as seis impressões digitais geradas para o texto de exemplo da figura 1. Para as impressões TF, FTF e FTMF, as *stop words*[4] são retiradas. A impressão TF é composta de apenas uma frase. A resolução não se aplica a este caso.

O texto teve três termos não encontrados no dicionário utilizado: “insurrecional”, “marcantes” e “republica”. Como a resolução foi definida como sendo de duas frases, a impressão digital FTI teve frases com os termos que apareceram mais acima do texto, uma vez que dois dos três termos possuem a mesma quantidade de caracteres. Para a impressão FEP, o resultado da equação apresentada na seção 2 definiu sua resolução como sendo 4, para a constante $k = 1$. Qualquer letra maiúscula encontrada é convertida para minúscula, antes mesmo da retirada da impressão digital.

[3] Trecho de texto sobre o movimento da inconfidência mineira, retirado em 06-10-2003, de http://www.geocities.com/athens/marathon/9563

[4] *Stop words* são palavras comuns da linguagem. Por este motivo, não representam bem o documento.

Tabela 1: Exemplo de impressão digital para as seis impressões definidas

Impressão digital	Exemplo	
1. TF	gerais minas produto 1789	
2. FTI	movimento insurrecional de 1789	característica marcantes que o
3. FEC	a inconfidência baiana de	em iniciar a revolta
4. FEP	das outras tentativas de	os mineiros que lideraram
	em minas gerais o	as demais capitanias o
5. FTF	minas gerais teve característica	minas gerais o plano
6. FTMF	teve característica marcantes que	a independência global do

3. Demais Etapas do Processo

3.1. Pesquisa e Coleta de Documentos Candidatos

A pesquisa utiliza um sistema de *metabusca* para a construção da base de documentos candidatos à similaridade. Cada frase da impressão digital do documento suspeito é usada como uma consulta simples em diversas máquinas de busca. O metabuscador consiste em um programa capaz de realizar consultas em máquinas de busca, podendo utilizar diferentes serviços de busca. Sua arquitetura é bem simples, uma vez que não precisa indexar documentos da Web, apenas consultá-los através dos serviços. MetaCrawler[5] e Miner[6] são exemplos de metabuscadores disponíveis na *Web*.

Para a realização deste trabalho desenvolvemos um metabuscador com arquitetura simplificada. O metabuscador simplesmente formata a consulta de acordo com o padrão utilizado pela máquina de busca TodoBr[7] e processa o resultado retornado, de forma a obter as URLs de resposta à consulta. Os documentos identificados por suas URLs podem ser recuperados e compor a base de documentos candidatos à similaridade.

3.2. Comparação Entre os Documentos

As etapas anteriores foram importantes para a construção da base de documentos candidatos à similaridade. A terceira etapa tem a função de comparar cada documento candidato com o documento suspeito, buscando verificar a similaridade entre os pares de documentos. Dois métodos foram utilizados: árvore Patricia e *Shingles*.

O primeiro método utiliza a árvore Patricia, estrutura de dados proposta em [Morrison, 1968]. A árvore Patricia é construída sobre o documento suspeito e os documentos candidatos têm seus conteúdos pesquisados na árvore, o que permite verificar a existência de longos trechos idênticos que ocorram no documento suspeito e em cada um dos candidatos. O segundo método utiliza o conceito de *shingles* [Broder, 1998] para medir a similaridade entre o documento suspeito e cada candidato, comparados dois a dois. Maiores informações sobre os métodos e algoritmos usados nesta etapa podem ser obtidas em [Pereira-Jr and Ziviani, 2003].

4. Resultados Experimentais

4.1. Construção de Coleções de Documentos Plagiados

Para a realização dos experimentos desenvolvemos um sistema gerador de documentos plagiados, utilizando trechos de documentos *Web*. O sistema foi desenvolvido de acordo

[5] http://www.metacrawler.com, 2004.
[6] http://www.miner.com.br, 2004.
[7] http://www.todobr.com.br, 2004.

com a intuição de que o usuário que utiliza a *Web* como fonte para a composição do seu documento não realiza, de forma significativa, alterações no texto plagiado. Assim, alterações como troca de palavras por sinônimos ou troca de termos de uma frase, mantendo o sentido original, não são tratadas pelo gerador, que simula uma *composição* de um documento a partir de outros documentos.

É necessário definir a quantidade de documentos *Web* que serão usados na composição do documento plagiado, bem como o tamanho, em número de termos, que o documento da composição deverá ter em relação ao tamanho dos documentos *Web* usados. O novo documento composto é chamado de documento "plagiado". O sistema inicialmente coleta os dez primeiros documentos retornados de consultas populares[8] realizadas na máquina de busca TodoBR. Em seguida é feita a leitura termo a termo do documento HTML para que seja retirado o texto, que é então separado em trechos (chamaremos de frases), definidos através de caracteres "ponto final". Frases aleatórias de cada documento são utilizadas na composição do documento plagiado, sempre mantendo o percentual de termos do documento candidato que está presente no documento plagiado.

4.2. Metodologia Utilizada nos Experimentos

Os experimentos foram realizados com o objetivo de verificar a capacidade do sistema em recuperar da *Web* o maior número possível de documentos que foram utilizados na composição do documento plagiado. Esses documentos vão compor a base de documentos candidatos. Os experimentos foram realizados buscando minimizar os custo do sistema que são: o número de requisições geradas na máquina de busca pela impressão digital, e o número de documentos que devem ser coletados para composição da base de documentos candidatos à similaridade. Assim, uma resolução maior, ou seja, que contém um número maior de frases, representa um custo maior para o sistema, uma vez que cada frase representa uma requisição à máquina de busca. Da mesma forma, coletar todos os documentos da resposta a uma consulta teria um custo maior que coletar somente o documento do topo do *ranking*.

Buscando reduzir o custo na realização dos experimentos, foi utilizada uma coleção reduzida de documentos plagiados no primeiro experimento, onde o melhor valor de granularidade é escolhido e usado nos próximos experimentos. Pelo mesmo motivo, a impressão digital de pior desempenho é excluída nos dois primeiros experimentos, e não mais utilizadas nos experimentos seguintes.

Os três experimentos para avaliação da etapa de geração da impressão digital tinham objetivos diferentes, mas foram realizados de forma semelhante, como mostra a figura 2. Inicialmente a impressão digital do documento plagiado é obtida. Em seguida cada frase da impressão é usada como uma consulta na máquina de busca TodoBR. As páginas retornadas pela consulta têm suas URLs comparadas com as URLs dos documentos usados na composição do documento plagiado, retornando então o percentual de documentos recuperados para aquela impressão digital.

4.3. Escolha do Melhor Valor para Granularidade

O primeiro experimento foi realizado com o objetivo de filtrar as impressões digitais utilizadas, escolhendo a melhor granularidade para cada impressão e excluindo aquela de pior resultado. Uma pequena coleção de 350 documentos plagiados foi utilizada. Com excessão das impressões FEP e TF, todas foram experimentadas com resoluções de 5, 10 e 15 frases e com granularidades de 4, 6 e 10 termos, combinando cada valor de resolução

[8] Utilizamos um arquivo de histórico diário de consultas do TodoBR, onde foram consideradas consultas realizadas de cinco a dez vezes no mesmo dia.

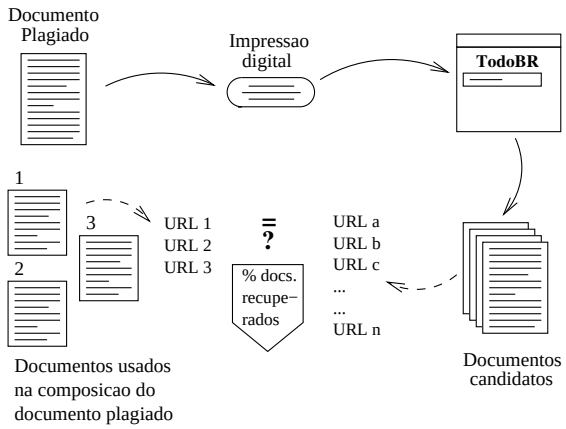


Figura 2: Modelo de experimento realizado para avaliar a etapa de geração da impressão digital

com os de granularidade. Para a impressão FEP, apenas os valores de granularidade foram variados, uma vez que a resolução é sempre definida por meio da equação apresentada na seção 2, neste caso com a constante $k = 2$. A resolução também não se aplica à impressão TF.

O gráfico da figura 3 faz a comparação entre os diferentes valores de granularidade, fazendo a média dos percentuais de documentos encontrados para as diferentes resoluções aplicadas. Percebemos que a maior granularidade experimentada, que foi de dez termos — o máximo permitido para consulta na maioria das máquinas de busca — apresentou os melhores resultados (exceto para TF), sendo este o valor de granularidade escolhido para os próximos experimentos. Para a impressão TF, formam consideradas 10, 30 e 50 páginas, sendo este último o de melhor resultado, como mostra a figura 3. Como esta impressão apresentou um baixo índice de documentos recuperados, ela será excluída dos próximos experimentos. Para as demais impressões digitais, os dez documentos do topo do *ranking* foram considerados.

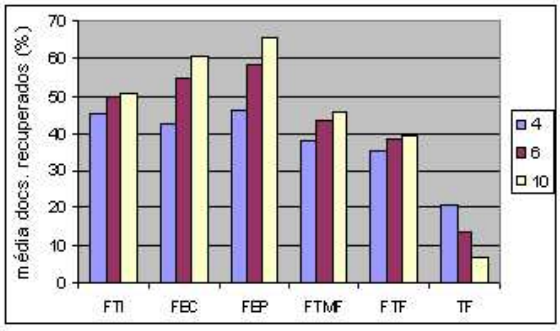


Figura 3: Comparação das diferentes granularidades para cada impressão digital

4.4. Impressões Digitais de Melhores Resultados

O experimento anterior foi útil para filtrar as possibilidades de impressões digitais para os documentos. Agora, temos o objetivo de avaliar a qualidade das impressões digitais para um número maior de documentos, tentando diminuir o custo para coleta. Foi utilizada uma coleção de 1.900 documentos plagiados para avaliar os resultados de cinco impressões digitais diferentes: FTI, FEC, FEP, FTF, FTMF, para três resoluções diferentes: 5, 10 e 15 frases. Para a FEP, a resolução é definida por meio da equação apresentada na seção 2, com dois valores para a constante k, que são $k = 1$ e $k = 2$, apresentando resoluções médias de $5,84$ e $12,15$ frases, respectivamente. A granularidade ficou fixada em dez termos, para todos as impressões digitais.

O gráfico da figura 4 faz uma comparação entre os percentuais médios de documentos recuperados, para cada impressão digital, com as diferentes resoluções. Impressões digitais de maior resolução apresentaram um melhor desempenho do que as impressões menores. Isto pode ser justificado pelo fato de que impressões maiores coletam um maior número de documentos.

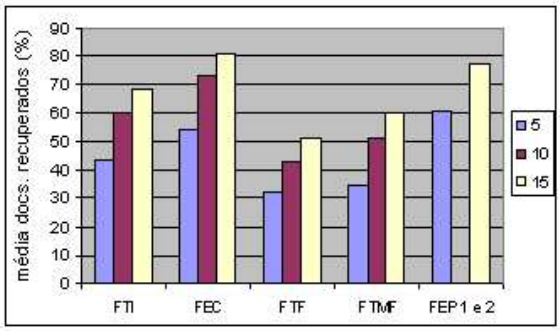


Figura 4: Comparação das diferentes resoluções para cada impressão digital

Na figura 4 vemos que a impressão de melhor resultado, a FEC com resolução 15, retornou $81,28\%$ dos documentos usados na composição do documento plagiado, seguido por FEP com $k = 2$, retornando $77,36\%$ dos documentos. A figura 5 apresenta o gráfico de pareto[9] para esta impressão, com valores acumulativos, classificando os índices de documentos recuperados de 10% em 10% (exceto para 100%). Verificamos que em $46,75\%$ dos casos, *todos* os documentos da composição foram recuperados. Somente em $8,71\%$ dos casos o desempenho ficou abaixo de 50%.

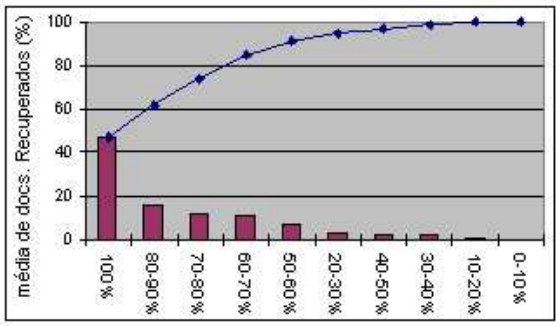


Figura 5: Gráfico de pareto para FEC com resolução 15.

Para este experimento, os *links* dos dez primeiros documentos retornados pelo sistema de busca foram analisados, em busca de algum documento que tenha sido usado na composição do documento plagiado. Coletar dez documentos de cada consulta realizada torna o processo caro em termos de coleta. Neste sentido, o experimento foi realizado de forma a também verificar qual era a posição do *ranking* do documento da composição encontrado. Verificamos que, em média, $81,66\%$ dos *documentos recuperados* estavam no topo do *ranking* e $93,12\%$ estavam ou no topo ou na segunda posição do *ranking*. Isto nos permite concluir que o desempenho do sistema, em termos de média de documentos recuperados, é pouco alterado quando se forma a base de documentos plagiados somente com os dois documentos do topo. O uso do sistema desta forma diminuiria o seu custo.

Fizemos uma análise manual buscando identificar situações específicas onde foi recuperado um número baixo de documentos usados na composição do documento plagiado, para a impressão FEC com resolução 15. Nessas situações vrificamos que os documentos usados na composição do documento plagiado eram: *home pages*, *blogs* com

[9] Gráfico de barras que enumera as categorias em ordem decrescente, da esquerda para a direita.

caracteres especiais, documentos contendo listas ou formulários. Estas verificações são indícios de que, nas situações em que um pequeno número de documentos foi recuperado, os documentos usados na composição do documento plagiado não eram boas representações de textos que normalmente são plagiados da *Web*. Assim, em uma situação real o sistema poderia apresentar melhor performance do que a verificada nos experimentos.

4.5. Combinação de Impressões Digitais

O experimento anterior buscou medir o desempenho do sistema para as diferentes impressões digitais de forma isolada. O objetivo agora é combinar as impressões, a fim de formar uma nova impressão com maior capacidade de recuperação de documentos similares. A mesma coleção do experimento anterior foi utilizada. A impressão de pior resultado do experimento anterior, FTF, não foi considerada. A resolução máxima considerada para as combinações foi de 30 frases. Desta forma, foi possível combinar todas as quatro impressões de resolução 5, ou combinar três a três as impressões com resolução de tamanho 10, ou ainda combinar duas a duas as impressões com resolução 15.

A nova impressão de melhor desempenho foi "FTI-FEC-FEP-10" (combinação das impressões FTI, FEC e FEP, com resolução 10 cada uma), seguida de "FTI-FEC-15", recuperando em média, respectivamente, $87,06\%$ e $86,63\%$ dos documentos usados na composição do documento plagiado. A figura 6 mostra o gráfico de pareto para a combinação "FTI-FEC-FEP-10". A análise do gráfico nos permite verificar um aumento significativo do desempenho para a nova impressão digital: em $61,53\%$ dos casos, *todos* os documentos da composição foram recuperados, contra $46,75\%$ da melhor impressão isolada, FEC, apresentada na figura 5. Isso representa um aumento de mais de 30% nas execuções que retornaram todos os documentos da composição, relacionado à impressão FEC. Para a mesma combinação, somente em $5,44\%$ dos casos o desempenho foi menor que 50%.

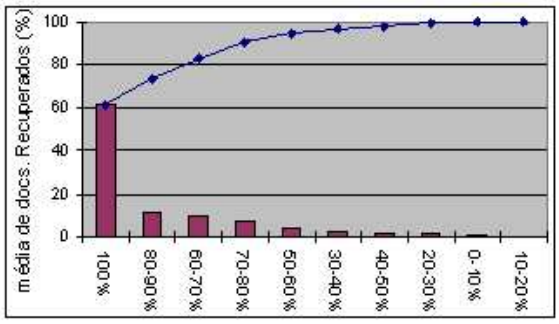


Figura 6: Gráfico de pareto para a combinação de impressões digitais "FTI-FEC-FEP-10".

5. Conclusões e Trabalhos Futuros

Um processo para detecção e recuperação de documentos similares na *Web* foi proposto e implementado. Através da construção de uma coleção de documentos plagiados, onde cada documento continha trechos de documentos da *Web*, foi possível medir e analisar o desempenho do processo.

O trabalho apresenta experimentos para medir o desempenho dos métodos utilizados na etapa de geração da impressão digital. Os experimentos foram realizados sobre uma coleção de documentos plagiados construída especialmente para este trabalho. Para a melhor impressão digital avaliada, em média 87% dos documentos usados na composição do documento suspeito são recuperados da *Web* e passam a compor a base de documentos

candidatos. Para a combinação de impressão digital "FTI-FEC-FEP-10", em quase 62% das execuções foi possível recuperar *todos* os documentos usados na composição do documento plagiado. Em média 93% destes documentos recuperados estavam entre os dois documentos do topo do *ranking*.

Como contribuições do trabalho, destacamos a proposta de um modelo eficaz para recuperação de documentos similares na *Web* e, ainda, um processo para avaliação do desempenho do modelo proposto, que pode ser utilizado para avaliar outros sistemas similares.

Uma sugestão de trabalho futuro é a construção de uma coleção de documentos plagiados a partir de documentos da *Web*, para ser disponibilizada para pesquisas em tópicos relacionados. As coleções utilizadas para este trabalho possuem tamanhos limitados (máximo de 1.900 documentos plagiados) e não estão estruturadas de forma a serem utilizadas com eficácia por terceiros. Para a construção desta coleção seria importante o levantamento estatístico do perfil de um documento plagiado. Com uma base de documentos que tenham sido manualmente alterados para fins de plágio, deve-se analisar os tipos de alterações que normalmente são feitas para, a partir daí, construir a coleção de documentos plagiados.

Referências

Brin, S., Davis, J., and Garcia-Molina, H. (1995). Copy detection mechanisms for digital documents. In *ACM SIGMOD Annual Conference*, pages 398–409, San Francisco.

Broder, A. (1998). On the resemblance and containment of documents. In *Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29. IEEE Computer Society.

Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1996). dscam : Finding document copies across multiple databases. In *4th International Conference on Parallel and Distributed Systems (PDIS'96)*, Miami Beach.

Garcia-Molina, H., Ketchpel, S. P., and Shivakumar, N. (1998). Safeguarding and charging for information on the internet. In *International Conference on Data Engineering (ICDE'98)*.

Manber, U. (1994). Finding similar files in a large file system. In *Proceedings of the USENIX Winter 1994 Technical Conference*, pages 1–10, San Fransisco, CA, USA.

Morrison, D. R. (1968). Practical algorithm to retrieve information coded in alphanumeric. *ACM*, 15(4):514–534.

Pereira-Jr, A. R. (2004). Recuperação de documentos similares na web. Master's thesis, Departamento de Ciência da Computação da Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

Pereira-Jr, A. R. and Ziviani, N. (2003). Syntactic similarity of web documents. In *First Latin American Web Congress*, pages 194–200, Santiago, Chile.

Shivakumar, N. and Garcia-Molina, H. (1995). Scam: A copy detection mechanism for digital documents. In *2nd International Conference in Theory and Practice of Digital Libraries (DL'95)*, Austin, Texas.

Stricherz, M. (2001). Many teachers ignore cheating, survey finds. *Education Week*. http://www.edweek.org/ew/ewstory.cfm?slug=34cheat.h20.

Proposta de uma Plataforma para Extração e Sumarização Automática de Informações em Ambiente Web

Carlos N. Silla Jr.*, Andre G. Hochuli*, Celso A. A. Kaestner

Pontifícia Universidade Católica do Paraná
Rua Imaculada Conceição 1155, 80215-901 Curitiba - PR - Brasil

{silla,hochuli,kaestner}@ppgia.pucpr.br

Resumo. *Neste trabalho é apresentada a arquitetura de uma plataforma automatizada para a extração de informações e sumarização de notícias em português, obtidas a partir da web. Esta plataforma foi aplicada a um estudo para a extração de informações sobre notícias de jogos de futebol. Foram realizados três experimentos visando estabelecer limites inferiores (baselines) para experimentos futuros, e também para observar o comportamento de sumarizadores existentes atuando de forma complementar à tarefa de extração.*

Abstract. *In this work we present the architecture of an automatic framework for information extraction and summarization of Brazilian Portuguese news, obtained from the web. This framework was applied to a case study for information extraction of news about soccer games. We performed three experiments with the goal of establishing baseline for future experiments and also to observe the behavior of the existing summarizers acting in a complementary way in the task of information extraction.*

1. Introdução

Com a explosão da Internet uma imensa massa de dados oriundos de diferentes fontes passou a se tornar disponível on-line. Um estudo recente de Berkeley [Lyman and H.R., 2003] mostra que em 2002 havia 5 milhões de terabytes de novas informações criadas em documentos impressos, filmes, mídias ópticas e magnéticas. A www sozinha contem cerca de 170 terabytes de informação, o que é equivalente a 17 vezes a coleção da biblioteca do congresso americano. Isto abriu aos usuários a oportunidade de se beneficiar destes dados sob muitas formas [Brin et al., 1998]. Em geral os usuários recuperam informações da Internet por meio de navegação nas páginas (*browsing*) ou pela busca direta de palavras-chave com auxílio de uma máquina de busca [Baeza-Yates and Ribeiro-Neto, 1999].

Entretanto estas estratégias de busca apresentam sérias limitações: (1) o processo de *browsing* não é adequado para a procura de itens de dados específicos, porque o seguimento de links é tedioso e facilmente o objetivo da busca é perdido; (2) a busca por palavras-chave, embora às vezes mais eficiente, retorna em geral grandes quantidades de dados, ultrapassando em muito as condições de manuseio do usuário.

Desta forma, apesar da sua disponibilidade e atualidade, os dados na Internet ainda não podem ser manipulados com tanta facilidade quanto às informações contidas em um Banco de Dados tradicional. Uma possível abordagem para o problema é a extração de dados das fontes disponíveis e sua transferência para uma representação estruturada, no

*Bolsistas PIBIC - CNPq/PUCPR

processo usualmente conhecido como Extração de Informações (*Information Extraction*) [Grishman, 1997].

As aplicações dos sistemas de Extração de Informações são muito variadas; como exemplo podem ser citadas: a obtenção dos autores e da data de apresentação de seminários [Freitag, 1998], a identificação de dados gerais sobre requisitos para empregados em uma base news de anúncios [Califf, 1998], a extração de informações financeiras a partir de sites Internet [Ciravegna, 2001], e a construção de um portal com sumarização e clipping das notícias mais importantes disponíveis nos sites de agências de notícias [McKeown et al., 2003]. Neste último caso também está envolvida a utilização de ferramentas para a sumarização de textos.

O processo de sumarização de textos envolve a construção de uma representação resumida do texto original - um sumário - que preserve as informações constantes no documento original, de acordo com as necessidades do usuário [Luhn, 1958]. Este assunto tem sido objeto de muitas pesquisas recentes [Sparck-Jones, 1999], [Mani, 2001], [Larocca Neto et al., 2002], [Pardo et al., 2003], [Silla Jr. et al., 2003], onde são empregadas técnicas oriundas do processamento de linguagem natural, da aprendizagem de máquina, e da análise e modelagem dos documentos. Pode-se dizer assim que a sumarização de textos atua de forma complementar à tarefa de Extração de Informações.

Neste trabalho é apresentada a arquitetura de uma plataforma automatizada para a extração de informações e sumarização de documentos. Essas tarefas podem ser realizadas independentemente uma da outra, ou pode-se utilizar a sumarização para, inicialmente, reduzir o tamanho do documento, e em seguida ser realizada a tarefa de extração de informações. A área de aplicação definida para os testes é a de extração e sumarização de notícias de futebol. Também são apresentados os resultados obtidos até o momento com a aplicação do sistema.

Este artigo está organizado da seguinte forma: a seção 2 apresenta a arquitetura proposta para o sistema; a seção 3 apresenta a metodologia dos testes efetuados e os resultados obtidos; e na seção 4 discutem-se as principais conclusões do trabalho e indicam-se as próximas etapas que serão realizadas no projeto.

2. A Arquitetura do Sistema

Uma visão geral da arquitetura da plataforma pode ser vista na Figura 1.

Inicialmente um robô de busca é utilizado para a coleta de páginas em sites pré-selecionados, neste trabalho são utilizados sites que contem notícias do campeonato paranense de futebol de 2004. Todas as notícias recuperadas são armazenadas em um banco de dados. Em seguida aplica-se um filtro para selecionar as notícias relevantes. Por exemplo, no contexto desta aplicação existem notícias extraídas da web que falam sobre outros assuntos envolvendo os times que participam do campeonato - como, por exemplo, a saída de um certo jogador para outro time ou futuros confrontos do time em questão - e que não envolvem partidas de futebol.

As notícias consideradas como relevantes que na aplicação descrevem partidas de futebol, vão passar por um pré-processamento, aonde serão extraídos os rótulos HTML. Com a coleção de documentos, pode-se utilizar tanto o processo de sumarização, como o de extração de informações. No caso da sumarização será gerado um sumário contendo as principais informações do texto, e este sumário pode ser apresentado como entrada para o extrator de informações. Já o extrator de informações é responsável por preencher um *template* pré-definido com as informações desejadas; no caso desejam-se informações sobre a partida em questão.

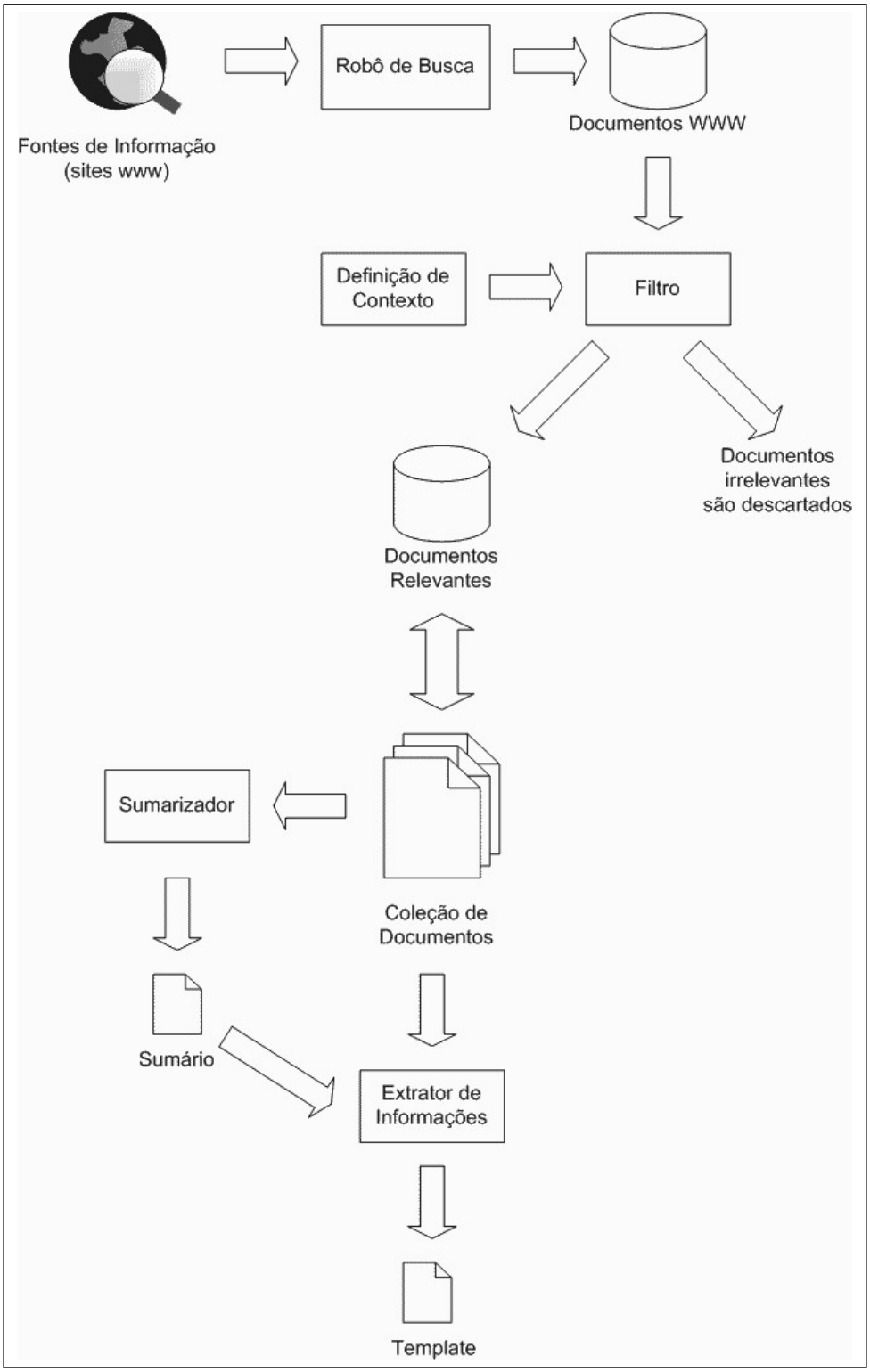


Figura 1: Visão Geral do Sistema

Para a recuperação de notícias dos sites, o robô de busca sendo utilizado é o WebSphinx[1], que foi desenvolvido em java e está muito bem documentado. Para cada fonte de notícia é customizado um robô específico, devido a grande diversidade na estrutura dos sites de notícias desse gênero.

A lógica utilizada pelo filtro é a seguinte: o filtro possui uma lista com todos os times do campeonato e seus possíveis apelidos, como por exemplo, São Paulo e Tricolor. Então, se a notícia em seu conteúdo possuir o nome de pelo menos dois times e também alguma forma de placar, como: [Digito a Digito] (3 a 1); [Digito x Digito] (1 x 0); ela é considerada como relevante.

Os documentos relevantes selecionados foram utilizados para o preenchimento de um *template*, cujas principais informações são: Time1, Time2, Placar. O preenchimento dos *templates* foi manual, por inspeção completa na base.

Em seguida aplicaram-se diversos sumarizadores (ver adiante) aos documentos a fim de verificar se o sumário gerado preserva a informação relevante para o preenchimento do *template*. Tanto o procedimento de sumarização quanto de extração de informações podem ser implementados usando algoritmos de aprendizado de máquina. No estado atual do projeto o método extração de informações usando algoritmos de aprendizado de máquina ainda está sendo implementado e por isso no momento o extrator está utilizado apenas um procedimento simples. O objetivo deste procedimento é o de se obter um valor de limiar mínimo (*baseline)* para os experimentos a serem realizados no futuro, visto que ele preenche o *template* com o time1 e o time2 com os times que mais apareceram na notícia respectivamente. Para isso o procedimento possui uma lista similar à existente no filtro para associar nomes de times e seus respectivos apelidos.

3. Testes Efetuados e Resultados Obtidos

Nesta seção são apresentados os resultados de três experimentos realizados com os seguintes objetivos: (1) verificar a performance do filtro; (2) verificar o acerto da identificação das informações consideradas indispensáveis: a identificação dos times pelo filtro; esta verificação visa utilizar estes resultados como *baseline* para os outros métodos que estão sendo implementados; (3) utilizar os sumarizadores de documentos anteriormente desenvolvidos no intuito de selecionar as sentenças mais relevantes dos documentos. Neste trabalho foram utilizados o FirstSentences Summarizer [Luhn, 1958], TF-ISF-Summ [Larocca Neto et al., 2000] e o ClassSumm [Larocca Neto et al., 2002].

Para a execução do projeto contruiu-se uma base de informações extraída diretamente da www e é composta de 1.169 notícias de sites que continham informações sobre o campeonato paranaense de futebol de 2004. A tabela 1 apresenta a matriz de confusão das notícias selecionadas pelo filtro. Dessa forma a Precisão do filtro na base é de 53,84% e o Recobrimento de 100%. O alto valor de recobrimento é devido a lógica usada pelo filtro, e mesmo tendo sua precisão em 53,84%, esse valor já reduz significativamente o número de notícias que deixaram de ser processadas. No intuito de verificar qual a real eficiência do filtro, são calculadas a micro e a macro médias, sendo que na micro-média considera-se a porcentagem total de acertos, e na macro-média considera-se inicialmente a porcentagem total de acertos por classe, para depois se efetuar a média dos valores obtidos para as classes. Conforme [Manning and Schutze, 2001], no primeiro caso procura-se ponderar a média por cada exemplo, enquanto que no segundo caso busca-se considerar eqüitativamente cada classe. Os resultados obtidos são: micro média de 93,33% e macro média de 96,38%.

[1]Disponível em: http://www-2.cs.cmu.edu/∼rcm/websphinx/

Tabela 1: Matriz de Confusão do Filtro

	Notícia Relevante	Notícia Não-Relevante
Relevante	91	78
Não-Relevante	0	1000

Como visto na seção 2, o extrator utiliza uma heurística para localizar quais foram os dois clubes que jogaram; apesar da tarefa parecer simples muitas vezes não o é, pois constatou-se que existem notícias na base que apresentam até mesmo seis times em seu conteúdo, comentando aspectos do jogo anterior ou do próximo confronto. Porém para se estabelecer um *baseline* para os experimentos futuros, foram verificadas as seguintes taxas de acerto: (1) somente com o nome do 1º time que jogou; (2) somente com o nome do 2º time que jogou; (3) com os nomes do primeiro e segundo times corretos; e por fim (4) os casos onde o nome dos times estivesse invertido. A tabela 2 apresenta os resultados obtidos, mostrando que mesmo uma técnica simples, que foi utilizada para estabelecer um *baseline*, obteve resultados de 48,51% de acerto para a identificação dos nomes dos times.

O terceiro experimento realizado procura verificar quão eficientes são alguns dos sumarizadores desenvolvidos anteriormente para atuarem de forma complementar a tarefa de extração de informação. Ou seja, qual a validade de se usar um sumarizador para tentar inicialmente selecionar as sentenças que contém a informação a ser extraída. Neste contexto um sumário ideal é aquele que possibilita preencher o *template* corretamente. Das notícias que compõem a base verificou-se que a maior parte delas possuem todas essas informações em uma única sentença; porém isso não acontece em todos os casos, podendo essa informação ser encontrada em duas ou até mesmo três sentenças.

Foram utilizados nos experimentos três sumarizadores: (1) FirstSentences que é um sumarizador normalmente utilizado como *baseline* em vários experimentos na área de sumarização; (2) TF-ISF-Summ (Term Frequency - Inverse Sentence Frequency Summarizer) que utiliza uma métrica adaptada do TF-IDF (Term Frequency - Inverse Document Frequency)[Salton et al., 1996] onde a noção de documento é substituída pela noção de sentenças; (3) ClassSumm que utiliza uma abordagem baseada em aprendizado de máquina, sendo que neste trabalho foi utilizado o algoritmo Naïve Bayes. Para realizar os experimentos, devido a necessidade de treinamento do ClassSumm, foi utilizado o método de validação cruzada com fator 10 (*ten-fold cross-validation*) [Mitchell, 1997].

Devido ao enfoque que está sendo utilizado para os testes, não basta comparar os resultados em termos de precisão e cobertura, e sim verificar se o resumo em questão permite preencher corretamente o *template* ou não. Foi estabelecido que, de cada sumarizador seriam escolhidas três sentenças, e que no caso do TF-ISF-Summ e ClassSumm, as sentenças apareceriam por ordem de importância e não pela ordem que estavam no texto (como é comumente utilizado nos experimentos da área de sumarização).

Dessa forma foi utilizado o seguinte método para comparar os três sumarizadores,

Tabela 2: Acerto do Extrator para Detectar os Times (%)

Critério	Número de casos corretos
Time 1	62,37
Time 2	48,51
Time 1 & Time 2	48,51
Time 1 & Time 2 invertidos	28,71

as tabelas 3, 4, 5 apresentam a porcentagem de acerto das sentenças selecionadas em relação a informação desejada, ou seja, quantas vezes cada sentença do sumário gerado contém o time1, o time2 e o placar.

Tabela 3: Acerto do Sumarizador First Sentences (%)

Método: First Sentences	Time1	Time2	Placar
Sentença 1	83,52	70,33	52,75
Sentença 2	14,29	24,18	34,07
Sentença 3	1,10	4,40	8,79
Total de Acerto	98,90	98,90	95,60
Não contém a informação	1,10	1,10	4,40

Tabela 4: Acerto do Sumarizador TF-ISF (%)

Método: TF-ISF-Summ	Time1	Time2	Placar
Sentença 1	70,33	59,34	32,97
Sentença 2	15,38	14,29	14,29
Sentença 3	9,89	10,99	12,09
Total de Acerto	95,60	84,62	59,34
Não contém a informação	4,40	15,38	40,66

Tabela 5: Acerto do Sumarizador ClassSumm (%)

Método: ClassSumm	Time1	Time2	Placar
Sentença 1	75,67	65,78	50,67
Sentença 2	18,78	22,11	28,67
Sentença 3	3,33	4,44	8,78
Total de Acerto	97,78	92,33	88,11
Não contém a informação	2,22	7,67	11,89

4. Conclusões e Direções Futuras

Os resultados apresentados neste trabalho são preliminares visto que o projeto ainda está em andamento. Verificou-se que realizar a etapa de sumarização antes da extração de informações auxilia significativamente o trabalho do extrator, uma vez que este terá que analisar apenas algumas sentenças para preencher o *template* ao invés de todo o documento.

No contexto desta aplicação, o sumarizador que possui o melhor desempenho é o FirstSencentes uma vez que este obteve um acerto de 98,90% para time1 e time2 e um erro de 4,40% para selecionar o placar. Enquanto que o TF-ISF-Summ obteve 95,60%; 84,62%; 59,34% e o ClassSumm 97,78%; 92,33%; 88,11%; de acertos para time1, time2 e placar respectivamente.

Conclui-se para que este tipo de notícia o uso das primeiras sentenças é indicado. Este resultado está em conformidade com a conjectura de que os resultados de um sumarizador dependem fundamentalmente da sua área de aplicação.

Como trabalho futuro será desenvolvido um sumarizador específico para o domínio que está sendo trabalhado, será implementado um extrator de informações utilizando algoritmos de aprendizado de máquina e serão realizados experimentos em outras áreas de aplicação.

Referências

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Brin, S., Motwani, R., Page, L., and Winograd, T. (1998). What can you do with a web in your pocket? *Data Engineering Bulletin*, 21(2):37–47.

Califf, M. E. (1998). Relational learning techniques for natural language extraction. Technical Report AI98-276, Univ. of Texas at Austion.

Ciravegna, F. (2001). Adaptive information extraction form text by rule induction and generalization. In *Proceedings of the 17 th. International Joint Conference on Artificial Intelligence, IJCAI'01*.

Freitag, D. (1998). Information extraction from HTML: Application of a general learning approach. In *Proceedings of the 15th. Conference on Artificial Intelligence, AAAI-98*, pages 517–523.

Grishman, R. (1997). Information extraction: Techniques and challenges. In *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, pages 10–27.

Larocca Neto, J., Freitas, A. A., and Kaestner, C. A. A. (2002). Automatic text summarization using a machine learning approach. In *XVI Brazilian Symposium on Artificial Intelligence*, number 2057 in Lecture Notes in Computer Science, pages 205–215, Porto de Galinhas, PE, Brazil.

Larocca Neto, J., Santos, A. D., Kaestner, C. A. A., and Freitas, A. A. (2000). Document clustering and text summarization. In *Proc. 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining (PADD-2000)*, pages 41–55, London: The Practical Application Company.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(92):159–165.

Lyman, P. and H.R., V. (2003). How much information. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 [Acesso em: 01/19/04].

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Company.

Manning, C. D. and Schutze, H. (2001). *Foundations of Statistical Natural Language Processing*. The MIT Press.

McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., Sable, C., Schiffman, B., and Sigelman, S. (2003). Projeto columbia newsblaster.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Pardo, T. A. S., Rino, L. H. M., and Nunes, M. G. V. (2003). Gistsumm: A summarization tool based on a new extractive method. In *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, number 2721 in Lecture Notes in Artificial Intelligence, pages 210–218, Germany.

Salton, G., Allan, J., and Singhal, A. (1996). Automatic text decomposition and structuring. *Information Processing and Management*, 32(2):127–138.

Silla Jr., C. N., Kaestner, C. A. A., and Freitas, A. A. (2003). A non-linear topic detection method for text summarization using wordnet. In *1º Workshop em Tecnologia da Informação e Linguagem Humana (TIL)*, São Carlos, SP, Brazil.

Sparck-Jones, K. (1999). *Advances in Automatic Text Summarization*, chapter Automatic Summarizing: factors and directions, pages 1 – 12. MIT Press.

Abducing Definite Descriptions Links

Sérgio A. A. Freitas[1] **, José Gabriel P. Lopes**[2] **, Crediné S. Menezes**[3]

[1]Departamento de Engenharia Elétrica, Universidade Federal do Espírito Santo,
Av. Fernando Ferrari, sn - 29600-090 - Vitória - ES, Brasil

[2]Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,
2825 - Monte da Caparica, Portugal

[3]Departamento de Informática, Universidade Federal do Espírito Santo,
Av. Fernando Ferrari, sn - 29600-090 - Vitória - ES, Brasil

sergio@inf.ufes.br,gpl@di.fct.unl.pt,credine@inf.ufes.br

Abstract. *In this article we propose a methodology to solve functional definite noun anaphora. Our approach uses an abductive scheme both to propose an antecedent and to find a possible relation between the anaphoric entity and its antecedent. The determination of the antecedent and the anaphoric relation uses among other informations: the gender, the number, the entity ontology and the focus in order to establish a set of pragmatic rules. Those pragmatic rules are used by the abductive scheme to solve the functional anaphora.*

1. Introduction

The resolution of an anaphoric definite noun phrase involves reasoning about the entities introduced by the discourse being interpreted by relating them with previously introduced entities. Take the following example [Sidner, 1979, pg. 42]:

(1) a. Horace got the picnic supplies out of the car.
b. The beer was warn.
b'. He had forgotten the beer.

Even though it is not explicitly conveyed, it is necessary to infer that there is a membership relation between the beer introduced in sentence (1b) and the picnic supplies in sentence (1a) (bridging phenomena [Heim, 1982]). It is also necessary to find an explanation for the definite noun phrases: *the picnic supplies* and *the car*, in the context where sentence (1a) occurs (assuming that there is no previous discourse).

Previous approaches to anaphora resolution namely: Focus Theory [Sidner, 1979], Centering [Grosz et al., 1995], and Carter's proposal [Carter, 1987], do not treat these phenomena (or they just treat them quite partially):

1. they assume an unique possible relation between the antecedent and the anaphoric particle: the *coref* relation. And this relation is not enough to treat phenomena such as the bridging occurring in sentence (1b): although the *picnic supplies* is the antecedent for the definite noun phrase *the beer*, the relation between these two entities should be: the beer is a member of the picnic supplies set.

2. when they try to analyse phenomena similar to bridging, they do not take in account the context influence on the anaphora resolution. In example (1), not all picnic supplies have beer in it and an automatic system should not waste time trying to predict all parts of a conveyed entity. We prefer to explain the relation of the most recently conveyed noun phrase in the context previously obtained by interpretation of previous sentences and beer is just an *optional* element that must be contextually related to the picnic supplies.

This led us to develop a methodology to solve definite noun anaphora (such as the bridging phenomena) in which the resolution process requires abductive inference of both: (1) the antecedent of an anaphoric particle, and (2) a plausible relation between them. For this we also build a discourse structure in order to organize the possible antecedents for an anaphora according to its salience (cf. [de Freitas and Lopes, 1994a]).

Previously, Hobbs et al [Hobbs et al., 1993] treated interpretation as an abduction problem. As a consequence, they solve some cases of definite reference, as a simple proof of their existence in a knowledge base, they lack to treat anaphoric phenomena where it is both needed to find an antecedent and a relation between the antecedent and the anaphoric particle. In this paper we concentrate in similar cases, proposing a methodology that abductively obtains both an antecedent and a plausible relation.

This paper is structured as follows: in section 2 we explain how a definite noun phrase is interpreted by our methodology. In section 3 we elaborate the abductive mechanism used for solving some cases of definite noun anaphora. In section 4 a non trivial example is used for illustrating the methodology previously introduced. In section 5 we discuss the usage of abduction to solve anaphora and, finally, in section 6, some conclusions are drawn.

2. Interpreting Definite Noun Phrases

Let $D = s_1, s_2, ..., s_{i-1}, s_i, s_n$ be a discourse and $s_1, s_2, ..., s_n$ its constituent sentences. The interpretation of a sentence s_i and of its definite noun phrases is achieved in two steps:

First, at the sentence level, a sentence s_i is translated into a DRS [Kamp and Reyle, 1993]. Indefinite noun phrases introduce discourse referents and adequate conditions on those referents and verbs introduce discourse conditions[1]. Definite noun phrases and null determiner noun phrases introduce discourse referents and the special condition **anchor(Atribs:Ref)**, which means that the referent *Ref* introduced by the noun phrase, needs to be anchored in the context provided by previously interpreted discourse.

Each entity is represented by a referent, *Ref*, and its associated attributes, $Atribs$. Entity attributes are a set of semantic pairs $< F, A >$, where F is a feature chosen among the set of 38 feature types for nouns (color, function, relation, place, size, etc.) presented by the Naive Semantics [Dahlgren et al., 1989], and A is the feature value. The appropriate features of a referent are the features that are most likely to express the semantic

[1] In this paper, we do not consider the referents introduced by the tense interpretation: the eventuality and the time interval referents [Rodrigues and Lopes, 1994].

attributes of the corresponding entity. For example, the features for a door (place, material, size, color) are different from those of a dog (name, owner, sex, size, color).

The resulting representation $DRS^i_{sentence}$ of a sentence s_i is a tuple: $< Refs_i, Conds_i, Anchored_i >$, where $Refs_i$ are the discourse referents and $Conds_i$ are the discourse conditions. Take the representation for each sentence in the following text:

(2) a. John bought a dog.
b. The animal barks a lot.

which is represented as:

$$DRS^{2a}_{sentence} =< \{X, Y\}, \{john(A : X), dog(B : Y), buy(A : X, B : Y)\} >$$
$$DRS^{2b}_{sentence} =< \{Z\}, \{animal(C : Z), anchor(C : Z), bark\_a\_lot(C : Z)\} >$$

During the second step, at the pragmatic interpretation level, each sentence semantic representation, $DRS^i_{sentence}$, is interpreted in the context $DRS^{i-1}_{discourse}$ provided by the interpretation of the previous $i-1$ sentences. All conditions **anchor(Atribs:Ref)** of $DRS^i_{sentence}$ must be abductively proved in the knowledge base[2] $Kb^{i-1}_{discourse}$. That is, given k conditions, $anchor^j_i(Atribs^i_j : Ref^i_j), j = 1, 2, ..., k$, of sentence s_i, one must abductively prove these conditions.

$$Kb^{i-1}_{discourse} \cup Ab_{anchor} \models \bigcup_{j=1}^{k} anchor^j_i(Atribs^i_j : Ref^i_j) \tag{3}$$

where Ab_{anchor} is the set of abduced literals that are necessary for proving those conditions and the abductive proof of the union of anchors requires the proof of each of the anchors.

As it will be further elaborated the abductible predicates are: **part_of, member_of, coref**.

If a condition $anchor^j_i(Atribs^i_j : Ref^i_j)$ of $DRS^i_{sentence}$can not be abductively proved in the context $Kb^{i-1}_{discourse}$ this is because either: (1) there is no previous discourse referent ***Ref*** that can be used as an antecedent, or (2) there is no suitable relation that can be abducted between $\textbf{\textit{Ref}}^i_j$ and its antecedent. In both cases, the entity denoted by $\textbf{\textit{Ref}}^i_j$ is assumed to be new information (like the referents introduced by an indefinite noun phrase [Kamp and Reyle, 1993, Heim, 1982]) and is simply accommodated in the discourse representation. This is represented by the following logic rules:

$$\begin{array}{ll} anchor(Atrib : Ref) & \leftarrow definite(Atrib : Ref). \\ anchor(Atrib : Ref) & \leftarrow indefinite(Atrib : Ref). \\ indefinite(Atrib : Ref) & \leftarrow not\, definite(Atrib : Ref). \end{array} \tag{4}$$

An entity, represented by referent ***Ref***, having semantic attributes ***Atrib***, is anchored in a discourse context if: (1) it can be proved to be a definite entity in this discourse, or (2) it is an indefinite entity in this discourse. An entity is indefinite if it can not be proved that it is definite. not is the default negation in Logic Programming.

[2]The set of all conditions of a discourse DRS is represented in a knowledge base Kb.

After the interpretation of $DRS^{i}_{sentence}$ in $Kb^{i-1}_{discourse}$ the resulting discourse representation will be:

$$DRS^{i}_{discourse} =< Refs_i, Conds_i, Anchored_i > \tag{5}$$

where $Refs = Refs(DRS^{i-1}_{discourse}) \cup Ref(DRS^{i}_{sentence})$, $Conds = Conds(DRS^{i-1}_{disc}) \cup Conds(DRS^{i}_{sentence}) \cup Ab_{anchor}$, and $Anchored$ is the list of referents whose anchor conditions were abductively proved to be definite in $Kb^{i-1}_{discourse}$.

An example is the interpretation of $DRS^{2b}_{sentence}$ in the context produced by $DRS^{2a}_{discourse} = DRS^{2a}_{sentence}$:

$$\begin{array}{rl} DRS^{2b}_{discourse} = & < \{X, Y, Z\}, \\ & \{john(A:X), dog(B:Y), buy(X,Y), animal(C:Z), \\ & bark\_a\_lot(Z), \mathbf{coref}(\mathbf{Z}, \mathbf{Y})\}, \\ & \{Z\} > \end{array} \tag{6}$$

The abduced relation $\mathbf{coref}(\mathbf{Z}, \mathbf{Y})$ is a plausible relation that could be assumed between the entity introduced by the definite noun phrase *the animal* and the previous introduced entity *dog*.

3. The abductive mechanism

Abduction [Kakas et al., 1992] is a form of reasoning in which given a set of sentences T (a theory presentation), a sentence G (an observation) and a set of sentences I (Integrity Constraints) the abductive task can be characterized as the problem of finding a set of sentences Ab (abductive explanation for G) such that:

1. $T \cup Ab \models G$,
2. $T \cup Ab$ satisfies I or $T \cup Ab \cup I$ is coherent

These rules characterize the Kowalsky's abductive scheme [Kakas et al., 1992] [Damásio et al., 1994], which is represented as the triple $< H, Ab, I >$, where H is a logic program such: $H \leftarrow L_1, ..., L_n$,where each L_i is either an atom A_i or its default negation $\sim A_i$, and H is an atom.

Abduction can be computed in a logic program H by extending the SLD and SLDNF resolution in such case that instead of failing in a proof when a selected subgoal fails to unify with the head of any rule, the subgoal can be viewed as a hypothesis. This is similar to viewing abducibles Ab as askable conditions which are treated as qualifications to answers to queries. As the set of abducibles are limited, the exponential problem presented by abduction inference can be reduced.

Kowalsky also proposes that a set of rules called Integrity Constraints I should be used to guarantee the coherence of the resulting knowledge base ($H \cup Ab \cup I$). The set of Integrity Constraints are user defined.

3.1. The Logic Program

In our approach, the logic program is the set of the conditions represented in $Kb^{i-1}_{discourse}$, the world knowledge about the entities and their semantic attributes, and the pragmatic rules that characterize the abductive mechanism (that will be showed in the following sections). $Kb^{i-1}_{discourse}$ represents the context in which the anchor conditions present at $DRS^{i}_{sentence}$ must be abductively proved.

3.2. The Set of Abductibles

We take as abducibles the following items:

member-of(Ref,Ref2), the entity denoted by referent *Ref* is a member of the entities denoted by *Ref2* if their attributes, $Atribs$ and $Atribs2$, respectively, are such that:
$features(Atribs) = features(Atribs2)$,
That is, the set of feature types for *Ref* is identical to the set of features types for *Ref2*, although their feature-values need not to be the same.
For example:

(7) a. The bus driver opened the doors.
b. The passengers used the back door to go out.
c. (and) the driver used the front door.

Here, both back door and front door can be members of a set of doors although they don't have the same value for the feature place.

coref(Ref,Ref2), the entity denoted by referent *Ref* corefers the entity denoted by *Ref2* if their associated semantic set, $Atribs$ and $Atribs2$, respectively, are such that:

1. $features(Atribs) \supseteq features(Atribs2)$,
2. $\forall F_1, F_2(F_1 \in features(Atribs) \& F_2 \in features(Atribs2) \& F_1 = F_2 \& \exists A_1, A_2 \mid A_1 = value(F_1) \& A_2 = value(F_2) \& A_1 = A_2)$.

For example:

(8) a. John bought a BMW.
b. The car arrived yesterday.

Both the car and BMW have have the same features, for example, *function.* Also the features' values are the same.

part-of(Ref,Ref2), the entity denoted by referent *Ref* is part of the entity denoted by *Ref2* as in:

(9) a. John bought a car.
b. The engine crashed yesterday.

Note that there is no need the engine's features to be directed related to the car's features. The classical solution will claim that there is a semantic network linking them. We claim that examples like (1) do not have a previous representation, instead there are clues the hearer uses to establish a link: the focus structure[de Freitas and Lopes, 1994b], the lexical information [de Freitas and Lopes, 1996], the usage of a definite article and a taxonomy of features.

3.3. The Integrity Constraints

The set of the integrity constraints used to test if $Kb_{discourse}$remains consistent after the abductive interpretation of $DRS^{i}_{sentence}$ in the context provided by $DRS^{i-1}_{discourse}$ are:

$$\begin{array}{l} \Leftarrow \text{member-of}(Ref, Ref2), \sim \text{member-of}(Ref, Ref2). \\ \Leftarrow \text{coref}(Ref, Ref2), \sim \text{coref}(Ref, Ref2). \\ \Leftarrow \text{part-of}(Ref, Ref2), \sim \text{part-of}(Ref, Ref2). \end{array} \tag{10}$$

It is not possible to state that a relation $\mathbb{R}$ (member-of, is-a, coref or part-of) between *Ref* and *Ref2* in $Kb_{discourse}$ and its negation hold at the same time for the same knowledge base. Also these relations can not be reflexive:

$$\begin{array}{l} \Leftarrow \text{member-of}(Ref, Ref2), \text{member-of}(Ref2, Ref). \\ \Leftarrow \text{part-of}(Ref, Ref2), \text{part-of}(Ref2, Ref). \end{array} \tag{11}$$

3.4. The mechanism

For the abductive mechanism an entity is anaphoric if it can be proved as definite in the previous discourse (as showed in (4)):

$$anchor(\mathcal{A} : Ref) \Longleftarrow definite(\mathcal{B} : Ref). \tag{12}$$

Where *Ref* denotes a referent with semantic attributes $\mathcal{A}$. This is done by finding a previous entity *Ref2* with semantic attributes $\mathcal{B}$, and by testing their attribute sets: $\mathcal{A}$ and $\mathcal{B}$ (features and values). Depending on the kind of relation Ψ (set identity, subset, superset) that can be coherently established between the attribute sets, a relation R between *Ref* and *Ref2* is abducted to complete such proof. This is summarized by the following rule:

$$\begin{array}{rl} definite(\mathcal{A} : Ref) \Longleftarrow & exists(\mathcal{B} : Ref2), \\ & \mathcal{A}\Psi\mathcal{B}, \\ & \mathcal{R}(\mathbf{Ref}, \mathbf{Ref2}). \end{array} \tag{13}$$

The following table describes the relations that could be abducted depending on how referents, *Ref* and *Ref2,* and their semantic attributes, $\mathcal{A}$ and $\mathcal{B}$, can be related:

	$value(A)$ valid-in $value(B)$	$value(A)$ not valid-in $value(B)$
$features(A) = features(B)$	member-of(*Ref,Ref2*) coref(*Ref,Ref2*)	member-of(*Ref,Ref2*)
$features(A) \subset features(B)$	part-of(*Ref, Ref2*)	—
$features(A) \supset features(B)$	coref(*Ref, Ref2*)	part-of(*Ref, Ref2*)

Tabela 1: Abductive Relations

Some notes about the above table:

1. $value(\mathcal{A})$ is valid in $value(\mathcal{B})$ iff for all features $F \in features(\mathcal{A}) \wedge F \in features(\mathcal{B})$, $value(\mathcal{A}) \in VV(\mathcal{B})$, where VV is the set of all valid values of a features F in $\mathcal{B}$. For example, if we try to relate a back door with a set of doors, the feature "place", although not instantiated, have valid features: back, front, middle, so the feature for the individual door, is still a valid one in the set of doors.
2. There are entries in the above table that have more that one possible abductible relation. This occurs because in some situations, it is not possible to predict which is the preferred relation, as in the example (1), where the entity *beer* could also be a $part\,of$ *the picnic supplies* (in this context). One could argue that this is not a valid relation: a part of relation must be established between entities that are constituents of each other. This is valid only if the entities and possible relations are previously represented at the world knowledge, what is not our case.
3. The part-of relation is very frequent. It expresses the idea in conversation that when we are sure that an object is related with another one, but we can not preciously state the relation, we assume that one is part of the other.

4. An example

Now we show how this works with the following example.

(14) a. A bus arrived at 5 pm.
b. The driver opened the doors.
c. The passengers used the back door to go out.

Lets consider the following representation for sentences (14a) and (14b):

$$DRS^{14a}_{sentence} = DRS^{14a}_{discourse}$$

X
bus(A:X)
arrive(A:X)

$$DRS^{14b}_{sentence} =$$

Y, Z
driver(B:Y), anchor(B:Y)
doors(C:Z), anchor(C:Z)
open(B:Y,C:Z)

In sentence (14b) the sentence interpretation of the definite noun phrases *the driver* and *the doors* introduce two anchor conditions and their respective semantic attributes $\mathcal{B}$ and $\mathcal{C}$. Now we must interpret $DRS^{14b}_{sentence}$ in the context provided by $DRS^{14a}_{discourse}$. This interpretation involves joining the referents of both DRSs, joining the common conditions of both DRSs, and finally accommodating each entity represented by the anchor conditions of $DRS^{14b}_{sentence}$. Such accommodation involves the abductive proof that both driver and doors are already definite in the previous knowledge base $Kb^{14a}_{discourse}$provided by the set of conditions of $DRS^{14a}_{discourse}$ plus the pragmatic rules presented in section (3).

The only possible antecedent for both is *the bus* introduced by sentence (14a). The resulting comparison between the semantic set for *the bus* ($\mathcal{A}$) and $\mathcal{B}$ and $\mathcal{C}$, are:

- they don't have the same feature set, and

- the features that are valid for $\mathcal{B}$ and $\mathcal{C}$ are not valid for $\mathcal{A}$.

So, the possible relation that could be abducted between $\mathcal{A}$ and $\mathcal{B}$, and $\mathcal{A}$ and $\mathcal{C}$ are the $part\,of$ relation. The result is the following representation:

$$DRS^{14b}_{discourse}= \boxed{\begin{array}{l} \text{X, Y, Z} \\ \text{bus(X),arrive(A:X)} \\ \text{driver(B:Y),part-of(X,Y)} \\ \text{doors(C:Z),part-of(X,Z)} \\ \text{open(B:Y,C:Z)} \end{array}} \{Y, Z\}$$

where {Y,Z} represent the elements that have been anchored, and could be used for future revision.

The interpretation of the sentence (14c) will lead the following representation:

$$DRS^{14c}_{sentence}= \boxed{\begin{array}{l} \text{W, S} \\ \text{passengers(D:W), anchor(D:W)} \\ \text{back\_ door(E:S), anchor(E:S)} \\ \text{get\_ out\_ through(D:W,E:S)} \end{array}}$$

Now we must abductively prove that the entities *passengers* and *back door* are already defined in the previous discourse $DRS^{14b}_{discourse}$. The semantic attributes of the passengers can not be related to the semantic attributes of the previous introduced entities, so the part-of relation is abduced to prove that the passengers are already defined in the discourse.

The attribute set of the entity back door $\mathcal{E}$ has the same set of features of the previously introduced entity, the doors $\mathcal{C}$, and all value for $\mathcal{E}$'s features are valid in $\mathcal{C}$'s features, so the $member\,of$ relation is abduced. The resulting representation is:

$$DRS^{14c}_{discourse}= \boxed{\begin{array}{l} \text{X, Y, Z, S, W} \\ \text{bus(X),arrive(A:X)} \\ \text{driver(B:Y),part-of(X,Y)} \\ \text{doors(C:Z),part-of(X,Z)} \\ \text{open(Y,Z)} \\ \text{passengers(D:W),part-of(X,Y)} \\ \text{back\_ door(E:S),member-of(S,Z)} \\ \text{get\_ out\_ through(D:W,E:S)} \end{array}} \{Y, Z, S, W\}$$

5. Using Abduction

Abduction was previously used as a tool to the natural language processing: temporal reasoning [Filho and de Freitas, 2003, Rodrigues, 1995] and discourse interpretation [Hobbs et al., 1993]. But why use it to solve anaphora?

We think that the answer is in Hobbs' et al article [Hobbs et al., 1993] where they propose an impressive integrated framework to solve some linguistic phenomena. Abduction is used to reason where there is incomplete information. Most of the human daily

reasoning use incomplete information. If someone observes the wet grass, the fact that it rained yesterday can be used as an explanation to the observation. Hobbs states this explanation as the new information the speaker transmits to the hearer.

We mostly agree with him, but also have two "observations": (1) in the daily reasoning, the explanation is only required if there is a need to it "if someone look at the wet grass and there is no need to reasoning about that, he could simply concluded that *the grass is wet*" and (2) Hobbs' usage of weighted abduction led to another problem "how weights can be determined?".

Kowalsky's abductive scheme [Kakas et al., 1992] implemented at Damasio, Alferes and Pereira's framework [Alferes and Pereira, 2002, Damásio et al., 1994] do not have these problems: only the observations that need explanation can be achieved and there is no weights.

This led us to our basic proposal: if an anaphora is observed there must be an explanation, which can be stated as "find an entity already introduced and link between this entity to the anaphora".

To the readers that want a more theoretical discussion about abduction (as a concept introduced by Charles Peirce) look at Deutscher's paper [Deutscher, 2002] where he discusses the (mis)use of abduction in linguistics.

6. Conclusion

We have shown that the abductive approach we have proposed for the definite noun anaphora resolution is a powerful methodology. It takes advantage of abductive mechanism and both determines a previous antecedent and a suitable relation with the anaphor. In this paper we have concentrated on the relations (part-of, member-of, coref), because the antecedents are determined using a focus structure which is not presented in this paper. The methodology incorporates both the pronominal anaphora and the definite nominal anaphora in the same framework.

The identified relations between the anaphoric entity and its antecedent are important because they give a more powerful interpretation to the anaphoric problem, enabling during a revision process to revise both a wrong antecedent or its relation with the anaphoric entity. We think that this relation revision process could be incorporated in a process for identifying and correcting misunderstandings (at discourse level) like the one proposed by McRoy and Hirst [McRoy and Hirst, 1995].

Although the set of possible relation that we use is small, we think it is sufficient to cover a large range of text.

References

Alferes, J. J. and Pereira, L. M. (2002). Logic programming updating - a guided approach. In A.Kakas and F.Sadri, editors, *Computational Logic: From Logic Programming into the Future - Essays in honour of Robert Kowalski*, volume 2 of *LNAI 2408*, pages 382–412. Springer-Verlag.

Carter, D. (1987). *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood Books.

Dahlgren, K., McDowell, J., and Edward P. Stabler, J. (1989). Knowledge representation for commonsense reasoning with text. *Computational Linguistics*, 15(3):149–170.

Damásio, C. V., Nejdl, W., and Pereira, L. M. (1994). Revise: An extended logic programming system for revising knowledge bases. In *Knowledge Representation and Reasoning*. Morgan Kaufmann.

de Freitas, S. A. A. and Lopes, J. G. P. (1994a). Discourse segmentation: Extending the centering theory. In *XI Simpósio Brasileiro de Inteligência Artificial*, UFCE - Fortaleza - CE.

de Freitas, S. A. A. and Lopes, J. G. P. (1994b). Improving centering to support discourse segmentation. In Bosch, P. and van der Sandt, R., editors, *Focus in Natural Language Processing*, volume 3 of *Working Papers of the Institute for Logic and Linguistics*. IBM, Heidelberg, Germany.

de Freitas, S. A. A. and Lopes, J. G. P. (1996). Solving the reference to mixable entities. In *Proceedings of the Indirect Anaphora Workshop*, University of Lancaster, Lancaster, UK.

Deutscher, G. (2002). On the misuse of the notion of 'abduction' in linguistics. *Journal of Linguistics*, 38:469–485.

Filho, A. M. C. and de Freitas, S. A. A. (2003). Interpretação do futuro do pretérito em narrativas. In *Anais do 1º workshop em Tecnologia da Informação e da Linguagem Humana, TIL'2003*, São Carlos - SP, Brasil.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2).

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachussets.

Hobbs, J. R., Stickel, M. E., Appelt, D. E., and Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Kakas, A., Kowalski, R., and Toni, F. (1992). Abdutive logic programming. *Journal of Logic Computational*, 2(6):719–770.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

McRoy, S. W. and Hirst, G. (1995). The repair of speech act misunderstandings by abductive inference. *Computational Linguistics*, 21(4):435–478.

Rodrigues, I. P. (1995). *Processamento de texto: Interpretação temporal*. PhD thesis, Universidade Nova de Lisboa.

Rodrigues, I. P. and Lopes, J. G. P. (1994). Temporal information retrieval from text. In Martin-Vide, C., editor, *Current Issues in Mathematical Linguistics*. North-Holland.

Sidner, C. L. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. PhD thesis, MIT, Cambridge, MA, USA.

HERMETO: A NL ANALYSIS ENVIRONMENT

Ronaldo Teixeira Martins[1,2], Ricardo Hasegawa[2], Maria das Graças Volpe Nunes[2,3]

[1]Faculdade de Filosofia, Letras e Educação – Universidade Presbiteriana Mackenzie
Rua da Consolação, 930 – 01302-907 – São Paulo – SP – Brazil

[2]Núcleo Interinstitucional de Lingüística Computacional (NILC) Av. Trabalhador São-Carlense 400 – 13560-970 – São Carlos – SP – Brazil

[3]Instituto de Ciências Matemáticas e da Computação — Universidade de São Paulo Av. Trabalhador São-Carlense 400 – 13560-970 – São Carlos – SP – Brazil

ronaldomartins@mackenzie.com.br, {rh,gracan}@icmc.usp.br

Abstract. *This paper describes HERMETO, a computational environment for fully-automatic, both syntactic and semantic, natural language analysis and understanding. HERMETO converts lists into networks and has been used to enconvert Brazilian Portuguese and English sentences into Universal Networking Language (UNL) hypergraphs.*

1. Introduction

The Universal Networking Language (UNL) [Uchida, Zhu and Della Senta, 1999; UNL Centre, 2003] is a knowledge-representation language expected to figure either as a pivot-language in multilingual machine translation systems or as a representation scheme in information retrieval applications. It has been developed since 1996, first by the Institute of Advanced Studies of the United Nations University, in Tokyo, and more recently by the UNDL Foundation, in Geneva, along with a large community of researchers - the so-called UNL Society - comprehending more than 15 different languages all over the world. As a semantic network, UNL is supposed to be logically precise, humanly readable and computationally tractable. In the UNL approach, information conveyed by natural language utterances is represented, sentence by sentence, as a hyper-graph made out of a set of directed binary labeled links (referred to as "relations") between nodes or hyper-nodes (the "Universal Words", or simply "UW"), which stand for concepts. UWs can also be annotated with attributes representing subjective, mainly deictic, information.

As a matter of example, the English sentence 'The sky was blue?!' would be represented in UNL as (1) below:

(1) aoj(blue(icl>color).@entry.@past.@interrogative.@exclamative, sky(icl>natural world))

In (1), 'aoj' is a relation (standing for 'thing with attribute'); 'blue(icl>color)' and 'sky(icl>natural world)' are UWs; and '@entry', `@past`, `@interrogative` and `@exclamative` are attributes. Differently from other semantic networks (such as conceptual graphs [Sowa, 1984, 2000] and RDF [Lassila and Swick, 1999]), UNL relations and attributes are pre-defined by the formalism. Relations constitute a fixed 44-relation set and

convey information on ontology structure (such as hyponym and synonym), on logic relations (such as conjunction and condition) and on semantic case (such as agent, object, instrument, etc) between UWs. The set of attributes, which is subject to increase, currently consists of 72 elements, and cope with speaker's focus (topic, emphasis, etc.), attitudes (interrogative, imperative, polite, etc.) and viewpoints (need, will, expectation, etc.) towards the event. In this sense, UNL is said to be able to represent not only denotative but also connotative, non-literal, information. The set of UWs, which is open, can be extended by the user, but any UW should be also registered and defined in the UNL Knowledge-Base (UNL KB) in order to be used in UNL declarations.

2. Enconverting from NL into UNL

Under the UNL Program, natural language analysis and understanding is referred to as a process of "enconverting" from natural language (NL) into UNL. The enconverting process is said to be not only a mere encoding (i.e., to rephrase the original sentence using different symbols) but actually to translate the source sentence in a new target language - the UNL -, which is supposed to be as autonomous and self-consistent as any NL, and whose graphs are expected to be language-independent and semantically self-governing.

In the UNL System, this enconverting process has been currently carried out either by the EnConverter (EnCo) [UNL Centre, 2002] or, more recently, by the Universal Parser (UP) [Uchida and Zhu, 2003], both provided by the UNL Center. In the first case, enconverting from NL to UNL is supposed to be conducted in a fully-automatic way, whereas in the second case a full-fledged human tagging of the input text should be carried out before NL analysis is performed. In both cases, results have not been that adequate. EnCo's grammar formalism, as well as UP's tagging needs, are rather low-level, and requires a human expertise seldom available. In what follows, we present an alternative analysis system, HERMETO, developed at the Interinstitutional Center for Computational Linguistics (NILC), in São Carlos, Brazil, which has been used for automatic enconverting from English and Brazilian Portuguese into UNL. Due to its interface debugging and editing facilities, along with its high-level syntactic and semantic grammar and its dictionary structure, it is claimed that HERMETO may provide a more user-friendly environment for the production of UNL expressions than EnCo and UP.

2. Motivations and Goals

HERMETO is a side product of two ongoing research and development projects carried out by NILC: POLICARPO and PUL∅. The former concerns the development of an English-to-Portuguese web translator, specialized in translating headlines and leads from the electronic edition of *The New York Times on the Web* into Brazilian Portuguese. PUL∅ concerns the development of a bimodal human-aided machine translation system for translating a Brazilian comics into LIST, a linearized version of Libras, the Brazilian Sign Language (for deaf people). Both systems are conceived as exclusively language-based, in the sense they are not supposed to require any extra-linguistic knowledge (as the one required in KBMT systems [Nirenburg et al, 1986]) neither a corpus of already translated samples (as in the case for EBMT systems [Furuse and Iida, 1992]). Additionally, both POLICARPO and PUL∅ were originally conceived as interlingua-based multilingual MT systems. Although the transfer approach might seem more suitable for each isolated task, our final goal is to provide a single system able to

process, bidirectionally, both the oral-auditive (English and Portuguese) and the sign-gesture (LIST) input and output.

UNL was chosen as the pivot language because of three main reasons: 1) it' s an electronic language for representing the semantic structure of utterances rather than its syntactic form; 2) the repertoire of UNL attributes can be extended to comprise semantic visual markers (as ' .@round' , ' .@square' , etc) required by sign language processing; and 3) as a multilingual and multilateral project, UNL could be used to assign cross-cultural interpretability to Portuguese and LIST texts.

In such a multilingual MT environment, HERMETO was conceived as an embedded NL analysis system, which should allow for developer' s customization and language parameterization. In its current state, it takes any plain text and enconverts it into UNL by means of a bilingual NL-UNL dictionary and a syntactic-semantic context-free grammar, both defined and provided by the user. The system was developed in C++ and is still bound to the Windows environment. HERMETO' s architecture is presented in the next section.

3. Architecture

HERMETO's architecture is presented in Figure 1 below. The input text - a plain text (.txt) written in ASCII characters - is split into sentences, each of which is tokenized and tagged according to the dictionary entries. Next, each sentence is traversed by a top-down left-to-right recursive parser, which searches for the best candidate matching as defined in the context-free grammar provided by the user. After parsing, the resulting synt actic structure is interpreted into UNL according to the projection rules written in the user's semantic grammar. The output is a UNL document, in its table form, i.e., as a list of binary relations embedded between UNL tags.

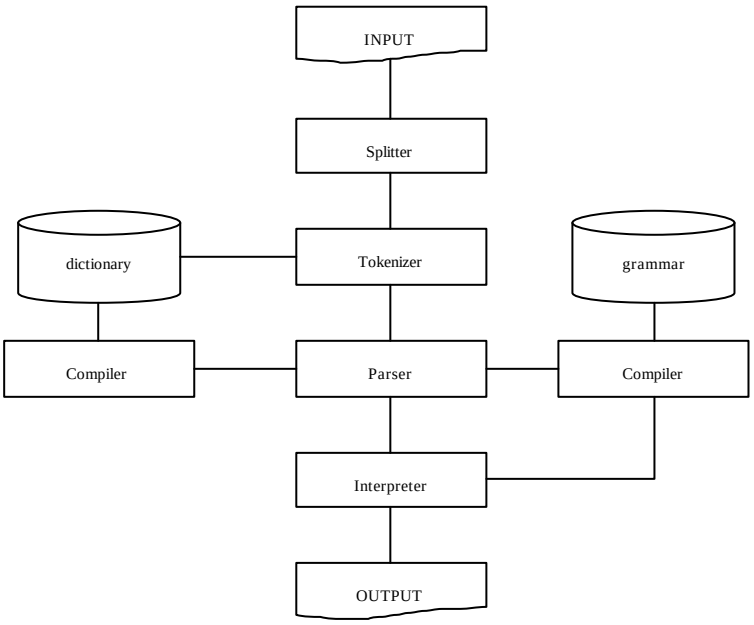


Figure 1. HERMETO's architecture

4. Resources

HERMETO's lingware consists of a bilingual NL-UNL dictionary and a NL-UNL transfer grammar. No other language resource (as the UNL KB, for instance) is required for the time being. Both dictionary and grammars are plain text files, which are automatically compiled by the machine. In order to improve grammar-writing tasks, HERMETO also comprises a grammar editor.

4.1 Dictionary

As EnCo, HERMETO takes a NL-UNL dictionary, whose entries, one per line, must be presented in the following format:

[NLE] {id} NLL "UW" (FEATURE LIST) <LG,F,P>;

NLE stands for "NL entry", which can be a word, a subword or a multiword expression, depending on the user's choice. NLL stands for "NL lemma". It is an optional field that can be used to clarify the string intended as NLE. The feature list consists of a list of attribute-value pairs, separated by comma. LG stands for a two-character language flag, according to the ISO 639. F and P indicate frequency and priority and are used for analysis and generation, respectively. Finally, any entry can be glossed and exemplified after the semi-colon.

The structure of HERMETO's dictionary is very much the same as EnCo's one: both dictionaries do not state any predefined structure, except for the syntax of each entry, and they can be customized by the user, who is supposed to decide the form of the entry, the need for lemmas and the set of attributes and the values they can take. However, there are three differences that should be stressed: 1) HERMETO compiles the plain text file itself, i.e., there is no need for a any extra compiling tool as DicBuild; 2) in HERMETO, the feature list is not a mere list of features but a list of attribute-value pairs, which allow for introducing variables in the grammar rules; and 3) HERMETO not only indexes but also compresses the dictionary (at the average rate of 65%).

Examples of dictionary entries are presented below:

[mesa] {} mesa "table(icl>furniture)" (pos:nou, gen:fem) <PT,1,1>;
[table] {} table "table(icl>furniture)" (pos:nou) <EN,1,1>;
[mesa] {} mesa "table(icl>furniture)" (pos:nou, ref:phy, fmt:squ) <LI,1,1>;

Except for the structure of the feature list and the language flag, HERMETO's dictionary formalism is the same as the one proposed in the EnCo's environment.

4.2 Grammar

HERMETO's grammar is a phrase-structure grammar defined by the 6-uple <N,T,P,I,W,S>, where N stands for the set of non-terminal symbols; T is the set of terminal symbols; P is the set of production rules; I is the set of interpretation rules; W is the weight (priority) of rules; and S stands for the start symbol. It is a context-free grammar, written in a plain text file, to be

automatically compiled by the machine. The set of terminal symbols to be used as variables should be defined in the top of the grammar file, and the mapping between this set and the dictionary attribute values should be stated at the end of the document.

The rules should follow the formalism: p –> i, where $p \in P$, and $i \in I$. P, which is the syntactic component, can be expanded as a[w] := b, where $a \in N$, $b \in N \cup T$, and $w \in W$. I, the semantic component, is expanded as a list of attributes and relations in the following format: $\mathbf{att_1, att_2, ..., att_n, rel_1, rel_2, ..., rel_n}$ where att stands for attributive rules, and rel stands for relational rules, both comprised in the UNL Specification.

Attributive and relational rules hold between positions (in the rule string) or indexes rather than words. The grammar also takes a given set of primitive operators (such as '[ ]', for optional; ' { }' , for exclusive; '< >' for lemma; '+' for blank space; '#' for word delimiter, etc.) in order to extend the expressive power of the formalism and reduce the necessary number of rules. The ' @entry' marker should be stated in every level, and the entry word is to be considered the head of each phrase. As in X-bar theory [6], entry word features are projected to and can be referred by the immediate higher level.

Examples of HERMETO's rules are presented below:

```
; 2.1.2. COMPLEX NOUN PHRASE (CNOP)
CNOP[2] := SNOP + 'and' + SNOP.@entry -> and(:03, :01)
CNOP[3] := SNOP + 'or' + SNOP.@entry -> or(:03, :01)
; 3.3. VERB
VERW[1] := ver.@entry - 'ied' -> :01.@past
VERW[1] := ver.@entry - 'ed' -> :01.@past
VERW[1] := ver.@entry - 'd' -> :01.@past
```

In such a grammar, context-sensitiveness can be stated as internal (dis)agreement between attribute values, such as in:

```
SNOP[1] := DET(GEN:x, NBR:y) + NOU(GEN:x, NBR:y).@entry -> :02.@def
```

The grammar is automatically compiled by HERMETO, which brings it to be an object-oriented scheme, where each non-terminal symbol is defined as an object, to be evoked by the others, during the syntactic and semantic processing. In order to optimize the compilation process, the length of each rule is limited to six symbols, and no nesting is admitted.

Although the expressive power of HERMETO's formalism may be the same as the one stated by EnCo, we claim that it is more intuitive, in the sense grammar writers are no longer supposed to be worried about the position of left and right analysis windows. They can work with (and even import) rules written according to more classic, high-level formalisms in NL understanding tradition.

5. Processes

HERMETO' s resources are parameters for more general, language-independent processes, as splitting, tokenizing, tagging, parsing and semantic processing. These constitute the NL analysis and UNL generation modules. In this sense, HERMETO can be seen as a

unidirectional transfer-based MT system itself, where NL is the source and the UNL is the target language.

5.1 Splitting, tokenizing and tagging

The process of sentence splitting, in HERMETO, is customized by the user, who is supposed to define, in the grammar, the intended set of sentence boundaries, such as punctuation marks and formatting markers, for instance. Each string of alphabetic characters or digits is considered a token, and blank spaces, as well as punctuation marks and non-alphabetic characters, are understood as word boundaries. Tagging is carried out through the dictionary, and no disambiguation decision is taken at this level. The word retrieval strategy seeks for the longest entries first, in the same way EnCo does. The word choice can be withdrawn, if HERMETO's parser comes to a dead-end situation.

5.2 Parsing

The tagged string of words is traversed by a chart parser, which applies the left (p) part of the grammar rules according to the priority defined by the user. Backtracking is supported, but cannot be induced. The parsing is rather deterministic, in the sense it provides only one parse tree for each sentence, the one best suited to the rules weight. Part-of-speech disambiguation is carried out during parsing, as the parser gets to the first possible parse tree. Parsing results can be exhibited by the interface and serve as the basis for semantic processing.

5.3 Semantic processing

Semantic processing is carried out together with parsing, in an interleaved way. Although semantic interpretation depends on the result of syntactic analysis, semantic projection rules are applied for any available partial tree, i.e., during the parsing itself. This does not cause, however, any parallelism between the syntactic and semantic modules, as the latter, although triggered by the former, cannot affect it. In this sense, HERMETO cannot deal with any generative semantics approach and is bound to the centrality of the syntactic component. Yet this can bring many difficulties in the UNL generation process, especially concerning the UW choice, i.e., word sense disambiguation, we have not advanced this issue more than EnCo does. The KB solution, which seems to be the most feasible one in EnCo environment, has not been adopted yet, for the trade-off still seems not to be positive, at least so far. As we have been mainly involved with an English sublanguage (the canned structure of English newspaper headlines and leads) and a regularized Portuguese (extracted from the comics), disambiguation can still be solved at the syntactic level.

6. Partial results

For the POLICARPO and the PUL∅ projects we have been working on the English-UNL and the Portuguese-UNL enconverting respectively. In the former case, we have compiled almost 1,500 web pages, downloaded in September 2002 from the *The NY Times* web site, to constitute our training and assessment *corpora*. Both English-UNL and UNL-Portuguese dictionaries have been already provided for every English word, except proper nouns, appearing in the corpus. The grammar has been split into a core grammar, common to every sentence, and five satellite grammars, specialized in 1) menu items, 2) headlines, 3) leads, 4)

advertisements and 5) others. Actually, we have observed that each of these sentence types convey quite different syntactic structures, which can be automatically filtered out of the general corpus. So far, we have already finished the core grammar and the one coping with menu items, and the precision and recall rates, for the assessment corpus, were 77% and 95% respectively, for complete UNL enconverting (i.e., UWs, relations and attributes). Althou gh menu items generally consists on quite simple single word labels, it should be stressed that many of them involved complex morphological structures that had to be addressed by the menu grammar. Anyway, HERMETO, together with the English-UNL dictionary and the core and menu grammars, has proved to be an interesting alternative for fully automatic English-UNL enconverting, at least in this case. For the time being, headlines have been already addressed, but no assessment has been carried out yet.

In PUL∅ project the coverage is rather small. Actually, the project is in its very beginning, and partial results concern a single story, for which HERMETO proved again, not only to be feasible for Portuguese-UNL enconverting, but to be easily integrated in a more complex system as well.

7. Shortcomings and further work

At the moment, we have been facing two main shortcomings: HERMETO accepts only ASCII codes and works only in Windows platform. Although we have planned to extend the current version to deal with Unicode and to run under other operational systems, we did not have the time to implement these changes. Furthermore, as we have been working rather on an English sublanguage (the NYT's one) and a sort of controlled (normalized) Portuguese, we have not really faced unrestricted NL analysis problems, which certainly will drive us to reconsider the UNL KB commitments. Therefore, in spite of the results achieved so far, HERMETO has still a long run before it can be considered a really feasible and suitable gener al NL-UNL enconverting environment. However, as former users of EnCo, we do believe it really represents a user-friendlier environment for fully automatic generation of UNL expressions out of NL sentences.

References

Furuse, O. and Iida, H. (1992), "Cooperation between transfer and analysis in example-based framework", In Proceedings of the 14th International Conference on Computational Linguistics, Nantes.

Lassila, O. and Swick, R. R. Resource Description Framework (RDF): model and syntax specification. W3C Recommendation, 1999.

Nirenburg, S, Raskin, V et al. (1986), "On knowledge-based machine translation', In Proceedings of the 11th International Conference on Computational Linguistics, Bonn.

Sowa, J. F., Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA, 1984.

Sowa, J. F., Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

Uchida, H. and Zhu, M. UNL annotation. Version 1.0. UNL Centre/UNDL Foundation, Geneva, 2003.

Uchida, H., Zhu, M. and Della Senta, T. A gift for a millennium, IAS/UNU, Tokyo, 1999.

UNL Centre. Enconverter specifications. Version 3.3. UNL Centre/UNDL Foundation, Geneva, 2002.

UNL Centre. UNL Specification. Version 3.2. UNL Centre/UNDL Foundation, Geneva, 2003

Impressões lingüísticas sobre duas axiomatizações para a Gramática Categorial

Luiz Arthur Pagani[1]

[1] Departamento de Lingüística, Letras Vernáculas e Clássicas – UFPR
Rua General Carneiro, n. 460, 11o. andar – 80.060–150 Curitiba – PR

arthur@ufpr.br

Abstract. *In the present essay, two different but logically equivalent axiomatizations of Categorial Grammar will be compared from an exclusively linguistic perspective: 1) the version of the so called reduction rules, and 2) the version of Lambek calculus. In order to do that, after a first section of introduction, each axiomatization will be separately presented in the second section. In the third section both versions will be commented in relation to the type of linguitic knowledge representation they allow. At the conclusion in the fourth section the reduction rules version is argued to be the best linguistic option. Although the observations are linguistically motivated, their conclusions affect the nature of the linguistic knowledge (especially the lexical one) to be represented in any parser for Categorial Grammar.*

Resumo. *No presente texto, compara-se de uma perspectiva exclusivamente lingüística duas axiomatizações diferentes, mas logicamente equivalentes, da Gramática Categorial: 1) a versão das chamadas regras de redução e 2) a do cálculo de Lambek. Para isso, depois de uma primeira seção de introdução, cada uma dessas axiomatizações é apresentada separadamente numa segunda seção. Na terceira seção, as duas versões são comentadas em relação ao tipo de representação do conhecimento lingüístico em cada uma delas. Por fim, na quarta seção, apresenta-se como conclusão a preferência pela axiomatização das regras de redução. E ainda que as observações sejam lingüisticamente motivadas, suas conclusões afetam a natureza do conhecimento lingüístico (especialmente o lexical) que precisa ser representado em analisadores para Gramáticas Categoriais.*

1. Introdução

No presente texto,[1] discutiremos de um ponto de vista exclusivamente lingüístico duas axiomatizações ligeiramente diferentes da Gramática Categorial, mas que são logicamente equivalentes. Numa dessas versões, a gramática é definida por um conjunto de seis pares de regras de redução (aplicação, permutação,[2] composição, promoção,[3] divisão

[1] Agradeço a meus companheiros do Laboratório de Lingüística, Lógica e Computação, da Universidade Federal do Paraná, por oferecer um ambiente propício ao estudo e à discussão de questões relacionadas à Gramática Categorial; agradeço em especial a Rodrigo Tadeu Gonçalves por algumas sugestões diretas sobre o presente texto.

[2] Normalmente, o termo usado aqui é "associatividade", do inglês *associativity*, mas seguindo uma observação de Oehrle, citada em [Wood, 1993, p. 37], prefiro usar o termo "permutação", traduzindo o inglês *swapping*, que remete à troca da ordem em que os argumentos se combinam com seu funtor.

[3] A escolha desse termo ainda é mais complicada do que a do anterior, porque em inglês mesmo o conceito é mencionado através de mais de um termo: *raising* [Wood, 1993, p. 42], *lifting* [Moortgat, 1988,

do funtor principal e divisão do funtor subordinado) que associa certas operações semânticas às respectivas operações de combinação categorial; na segunda versão, essas mesmas regras são teoremas deriváveis a partir de um único axioma e de dois pares de regras de inferência, ou dedução — um par de regras para cada conectivo categorial (/ e \).[4]

Nesta apresentação, no entanto, não nos deteremos no aspecto lógico ou dedutivo das axiomatizações. O principal objetivo aqui não é discutir a decidibilidade ou a precisão axiomática de cada um dos dois sistemas, mas sim chamar a atenção para certos aspectos ontológicos e epistemológicos da representação do conhecimento lingüístico. Para isso, ao invés de postularmos alguma espécie de primazia algébrica ou computacional de um sistema em relação ao outro (que talvez nem exista, se eles forem mesmo logicamente equivalentes), vamos decidir essa primazia a partir de determinadas características que cada um desses sistemas apresenta em relação ao tipo de explicação que ele pode sugerir para o trabalho de um lingüista.

2. Gramáticas Categoriais

Segundo [Moortgat, 1988, pp. 1–2], uma Gramática Categorial se distingue de outras teorias lingüísticas muito semelhantes (tais como a Gramática de Estrutura Sintagmática Generalizada (*Generalized Phrase Structure Grammar*, GPSG) ou a Gramática de Estrutura Sintagmática Conduzida pelo Núcleo (*Head-Driven Phrase Structure Grammar*, HPSG)) por apresentar as seguintes quatro características:

- *Lexicalismo*. As teorias gramaticais que se concentram na estrutura superficial compartilham uma tendência em deslocar para o léxico a carga explicativa que, em outras teorias, seria atribuída ao componente sintático. Ao desenvolver uma noção mais ampla de estrutura categorial, por exemplo, a GPSG torna desnecessário um componente transformacional como o da Gramática Gerativa Clássica. A Gramática Categorial avança um passo a mais em direção ao lexicalismo, tornando desnecessário o próprio componente sintagmático. A informação sintática é completamente projetada a partir das estruturas categoriais atribuídas aos itens lexicais. Na sua forma mais pura, a Gramática Categorial identifica o léxico como o único local para as estipulações específicas às línguas. A sintaxe é uma álgebra livre: uma combinatória universal conduzida pelas estruturas categoriais complexas.
- *Estrutura de função e argumento*. A contribuição categorial mais específica para a teoria das categorias é a de que as expressões incompletas são modeladas, sintática e semanticamente, como funtores. As dependências elementares entre as expressões, que determinam fenômenos como a regência, o controle e a concordância, são todas definidas através da hierarquia entre funções e argumentos, e não por sua configuração estrutural.
- *Flexibilidade dos constituintes*. A Gramática Categorial Clássica, assim como a Gramática de Estrutura Sintagmática, atribui uma única estrutura de constituin-

p. 11] e *shifting* [Carpenter, 1997, p. 100]. Preferimos então o termo "promoção", mas em [Neto, 1999] o termo usado é "elevação"; só achamos inadequado o uso de "alçamento", porque esse termo tem sido usado na Gramática Gerativa para designar um outro tipo de fenômeno (o movimento de um constituinte para uma posição mais alta na árvore de estrutura sintagmática da expressão).

[4]Na verdade, no cálculo de Lambek há um terceiro conectivo (•), o que exige mais um par de regras de inferência; no entanto, a ausência desse terceiro conectivo e de suas respectivas regras de dedução não afetam as questões discutidas no presente texto. Além disso, ambas as axiomatizações são compostas efetivamente por esquemas de axiomas e de regras de inferência, que ainda precisam ser preenchidos por categorias para se tornarem axiomas e regras; no entanto, tomaremos aqui a liberdade terminológica de chamá-los apenas de axiomas e de regras de inferência.

tes a uma expressão não-ambígua. As teorias categoriais generalizadas substituem essa noção de constituintes por outra mais flexível, oferecendo um inventário mais amplo de operações combinatórias que configuram um cálculo da mudança de tipo. Uma expressão não-ambígua é associada a um conjunto de derivações equivalentes. A coordenação booleana generalizada funciona como uma técnica experimental que revela os constituintes alternativos ocultos.

- *Composicionalidade*. A relação entre a álgebra sintática e a álgebra semântica é um homomorfismo, ou seja, uma relação que preserva a estrutura, na qual cada operação sintática corresponde a uma operação semântica. A Gramática Categorial Clássica incorpora uma forma de composicionalismo bastante forte, baseada na correspondência entre a regra de redução sintática central e a aplicação funcional na semântica. Os sistemas categoriais generalizados ampliam esta forma forte de composicionalidade para o cálculo da mudança de tipo, executando assim o programa da interpretação conduzida pelo tipo.

Uma conseqüência dessas características é que muitos dos fenômenos lingüísticos que exigiram da Gramática Gerativa, por exemplo, a postulação de categoria vazia, deslocamento e eliminação de estrutura, poderão ser monotonicamente resolvidos na Gramática Categorial sem recorrer a operações destrutivas como estas.

2.1. Regras de Redução

Numa de suas versões mais difundidas entre os poucos lingüistas que adotam a Gramática Categorial, normalmente ela é definida através de um conjunto de seis pares de regras de redução[5] que associam uma operação de combinação categorial e uma operação de construção da representação semântica.

Nessa versão, seguindo [Moortgat, 1988, p. 11],[6] os seis pares de regras de redução são definidos da seguinte maneira:

Regras de redução

R1 Aplicação

$X/Y : f,\ Y : a \Rightarrow X : f(a)$

$Y : a,\ Y\backslash X : f \Rightarrow X : f(a)$

R2 Composição

$X/Y : f,\ Y/Z : g \Rightarrow X/Z : \lambda v[f(g(v))]$

$Z\backslash Y : g,\ Y\backslash X : f \Rightarrow Z\backslash X : \lambda v[f(g(v))]$

R3 Permutação

$(Z\backslash X)/Y : f \Rightarrow Z\backslash(X/Y) : \lambda v_1[\lambda v_2[f(v_2)(v_1)]]$

$Z\backslash(X/Y) : f \Rightarrow (Z\backslash X)/Y : \lambda v_1[\lambda v_2[f(v_2)(v_1)]]$

R4 Promoção

[5] Na verdade, esta é a versão mais ampla dessa vertente. A primeira versão da Gramática Categorial, proposta por [Ajdukiewicz, 1935] tinha apenas uma das regras de aplicação, porque ela não era direcional; numa das primeiras aplicações mais lingüisticamente motivadas [Bar-Hillel, 1953], conhecida como modelo AB, apenas o par de regras de aplicação era usado; numa outra versão, conhecida como Gramática Categorial Livre [Cohen, 1967], além das regras de aplicação, aparecem também as regras de permutação, de composição e de promoção. No entanto, o modelo mais empregado modernamente é mesmo esse de seis pares de regras, que também aparece num formato um pouco alterado na formulação de [Steedman, 1988], chamada de Gramática Categorial Combinatória.

[6] As expressões do cálculo lambda que representam a interpretação semântica receberam uma notação um pouco mais explícita aqui. No texto original, o escopo do operador lambda era marcado com um ponto, como em $\lambda P.P(x)$, o que é bastante usual; no entanto, como em fórmulas muito longas pode ficar difícil perceber o escopo do operador, preferi uma notação na qual o escopo recebe um marcador de início e de fim, como em $\lambda P[P(x)]$ (onde os colchetes marcam inequivocamente o começo e o final do escopo do operador).

$X : a \Rightarrow Y/(X\backslash Y) : \lambda v[v(a)]$
$X : a \Rightarrow (Y/X)\backslash Y : \lambda v[v(a)]$

R5 Divisão (funtor principal)

$X/Y : f \Rightarrow (X/Z)/(Y/Z) : \lambda v_1[\lambda v_2[f(v_1(v_2))]]$
$Y\backslash X : f \Rightarrow (Z\backslash Y)\backslash(Z\backslash X) : \lambda v_1[\lambda v_2[f(v_1(v_2))]]$

R6 Divisão (funtor subordinado)

$X/Y : f \Rightarrow (Z/X)\backslash(Z/Y) : \lambda v_1[\lambda v_2[v_1(f(v_2))]]$
$Y\backslash X : f \Rightarrow (Y\backslash Z)/(X\backslash Z) : \lambda v_1[\lambda v_2[v_1(f(v_2))]]$

De acordo com estas regras, e considerando que as expressões "Pedro", "ama" e "Maria" correspondem respectivamente aos pares de categoria sintática e representação semântica '$N : p$', '$(N\backslash S)/N : A$' e '$N : m$', podemos representar a estrutura da sentença "Pedro ama Maria" através de um diagrama como o da Figura 1.[7]

$$\dfrac{\dfrac{\text{Pedro}}{N : p}\,Lx \quad \dfrac{\dfrac{\text{ama}}{(N\backslash S)/N : A}\,Lx \quad \dfrac{\text{Maria}}{N : m}\,Lx}{N\backslash S : A(m)}\,R1}{S : A(m)(p)}\,R1$$

Figura 1: Derivação de "Pedro ama Maria" apenas com R1

Esse diagrama da Figura 1 representa a demonstração de que a sentença "Pedro ama Maria" é uma sentença que denota a relação de amar que se estabelece de Pedro para Maria ('$S : A(m)(p)$'), a partir das regras de redução e de três premissas: 1) "Pedro" é um nome que denota o indivíduo Pedro ('$N : p$'), 2) "ama" é um predicado de dois lugares que denota a relação de amar ('$(N\backslash S)/N : A$') e 3) "Maria" é um nome que denota o indivíduo Maria ('$N : m$').

Nesse sentido, esse diagrama não é muito diferente de um diagrama em árvore, como os que são associados a uma Gramática de Estrutura Sintagmática. A grande diferença, no entanto, é que, a partir de uma Gramática de Estrutura Sintagmática para "Pedro ama Maria", consequiríamos construir uma única árvore para essa sentença; já com uma Gramática Categorial, como a apresentada acima, poderíamos chegar a uma outra estrutura para a mesma sentença, chegando à mesma representação semântica final, através de uma seqüência de regras diferente da anterior, como podemos ver no diagrama da Figura 2.[8]

Poderíamos chegar ainda a um terceiro diagrama para a mesma sentença "Pedro ama Maria", como na Figura 3, também com a mesma representação semântica. Na verdade, na Gramática Categorial, podemos encontrar um número infinito de derivações equivalentes para a mesma expressão, o que normalmente é chamado de ambigüidade espúria, que já foi considerado uma das principais falhas da Gramática Categorial, mas que pode ser facilmente controlada através de uma exigência de normalização das derivações (como em [Carpenter, 1997, pp. 160–164]).

[7] Nestes diagramas, conhecidos como dedução ao estilo de Prawitz, as barras horizontais relacionam uma conclusão e suas premissas, de forma que a conclusão aparece debaixo da barra e as premissas sobre ela; ao lado direito da barra registra-se a regra empregada na inferência. No entanto, a inserção dos itens lexicais não segue bem esse padrão: as expressões lingüísticas aparecem acima das suas respectivas barras na primeira linha do diagrama, e debaixo delas são registradas suas respectivas categorias e representações semânticas. Finalmente, algumas derivações exigem ainda a introdução de suposições, que são apresentadas entre colchetes numerados com um índice, que marca o escopo entre a sua introdução e a sua eliminação.

[8] A redução-β é uma das principais operações do cálculo-λ, e pode ser caracterizada pela seguinte fórmula: $\lambda v[F](a) \Rightarrow F[v \mapsto a]$, que pode ser lida como 'um termo-λ ($\lambda v[F]$), aplicado a outro termo (a) é equivalente ao termo no escopo do operador com as ocorrências livres da variável v substituídas pelo termo a ($F[v \mapsto a]$)' [Carpenter, 1997, p. 50].

$$\dfrac{\dfrac{\dfrac{\dfrac{\text{Pedro}}{N:p}\,Lx \quad \dfrac{\text{ama}}{(N\backslash S)/N:A}\,Lx \quad \dfrac{\text{Maria}}{N:m}\,Lx}{N\backslash(S/N):\lambda x[\lambda y[A(y)(x)]]}\,R3}{S/N:\ \lambda x[\lambda y[A(y)(x)]](p) =_{red.\beta} \lambda y[A(y)(p)]}\,R1}{S:\ \lambda y[A(y)(p)](m) =_{red.\beta} A(m)(p)}\,R1$$

Figura 2: Derivação de "Pedro ama Maria" com R1 e R3

$$\dfrac{\dfrac{\dfrac{\dfrac{\text{Pedro}}{N:p}\,Lx \quad \dfrac{\text{ama}}{(N\backslash S)/N:A}\,Lx \quad \dfrac{\text{Maria}}{N:m}\,Lx}{S/(N\backslash S):\lambda P[P(p)]}\,R4}{S/N:\ \lambda x[\lambda P[P(p)](A(x))] =_{red.\beta} \lambda x[A(x)(p)]}\,R2}{S:\ \lambda x[A(x)(p)](m) =_{red.\beta} A(m)(p)}\,R1$$

Figura 3: Derivação de "Pedro ama Maria" com R1, R2 e R4

No entanto, do ponto de vista lingüítico, pode-se perceber uma pequena diferença entre o diagrama da Figura 1, por um lado, e os diagramas das Figuras 2 e 3, por outro: se considerarmos que as operações realizadas pelas regras de redução equivalem também a concatenações das seqüências fonológicas, na Figura 1 teríamos uma estrutura prosódica correspondente a "(Pedro (ama Maria))" (onde os parênteses encerram os constituintes prosódicos concatenados), enquanto que nas Figuras 2 e 3 teríamos a seguinte estrutura prosódica: '((Pedro ama) Maria)'.

Apesar de semanticamente equivalentes, essas duas estruturas prosódicas apresentam características sintáticas e discursivas distintas. Por exemplo, apenas a segunda estrutura aceitaria uma continuação como "E não Márcia", no sentido de que não é a Márcia, e sim a Maria, a pessoa que o Pedro ama; a primeira estrutura prosódica, ao contrário, é compatível com uma continuação como "E não Paulo", no sentido de que a pessoa que ama a Maria não é o Paulo, e sim o Pedro. Isso justifica, do ponto de vista lingüístico, a distinção das infinitas derivações aparentemente equivalentes de "Pedro ama Maria" em dois grupos: 1) o das que concatenam primeiro "Pedro" e "ama", e 2) o das que concatenam primeiro "ama" e "Maria".

2.2. Cálculo de Lambek

Ainda segundo [Moortgat, 1988, p. 2],

> o cálculo de Lambek substitui o conjunto de regras de redução categorial que foram propostas na literatura (Aplicação, Composição, Promoção, etc.) por uma noção geral de derivabilidade, a partir da qual as leis de redução são consideradas teoremas. A derivabilidade é definida na forma de axiomatização de seqüentes, o que reduz as derivações categoriais a deduções lógicas com base nos procedimentos de prova desenvolvidos originalmente por Gentzen, em seu trabalho sobre o cálculo proposicional intuicionístico.

Devido à facilidade notacional, ao invés dos diagramas de derivação de seqüentes, vamos adotar aqui a mesma notação de derivação da Dedução Natural usada nos diagramas anteriores para a sentença “Pedro ama Maria”. Assim, ao invés de seis pares de regras, no cálculo de Lambek, precisamos apenas de um par de regras para cada conectivo: um de introdução e outro de eliminação do conectivo, como nos esquemas da Figura 4, abaixo, adaptados de [Carpenter, 1997, pp. 153 e 156].

$$\frac{X/Y : f \quad Y : a}{X : f(a)}\ /E \qquad \frac{Y : a \quad Y\backslash X : f}{X : f(a)}\ \backslash E$$

(a) Eliminação de / (b) Eliminação de \

$$\frac{\begin{array}{c}\vdots \quad [Y : v]^n \\ \vdots \quad\ \vdots \end{array}}{\dfrac{X : F}{X/Y : \lambda v[F]}\ /I^n} \qquad \frac{\begin{array}{c}[Y : v]^n \quad \vdots \\ \vdots \quad\ \vdots \end{array}}{\dfrac{X : F}{Y\backslash X : \lambda v[F]}\ \backslash I^n}$$

(c) Introdução de / (d) Introdução de \

Figura 4: Esquemas para eliminação e introdução dos conectivos

Assim, nesta outra axiomatização, “as leis de redução que foram introduzidas antes como primitivos, passam a ser *teoremas*; ou seja, inferências válidas da lógica dos conectivos categoriais” [Moortgat, 1988, p. 27].

Em relação ao diagrama da Figura 1, com a derivação da sentença “Pedro ama Maria” apenas com a regra R1, a derivação empregando apenas os esquemas de eliminação dos conectivos, na Figura 5, não apresenta nenhuma diferença.

$$\dfrac{\dfrac{\text{Pedro}}{N : p}\ Lx \qquad \dfrac{\dfrac{\text{ama}}{(N\backslash S)/N : A}\ Lx \qquad \dfrac{\text{Maria}}{N : m}\ Lx}{A(m)}\ /E}{A(m)(p)}\ \backslash E$$

Figura 5: Derivação de “Pedro ama Maria” apenas com eliminação

No entanto, o diagrama equivalente à derivação com R1 e R3 apresenta uma diferença essencial, que é a suposição de uma variável da categoria ‘N’, que aparece entre colchetes na derivação da Figura 6.

$$\dfrac{\dfrac{\dfrac{\dfrac{\text{Pedro}}{N : p}\ Lx \qquad \dfrac{\dfrac{\text{ama}}{(N\backslash S)/N : A}\ Lx \qquad [N : x]^1}{N\backslash S : A(x)}\ /E}{S : A(x)(p)}\ \backslash E}{S/N : \lambda x[A(x)(p)]}\ /I^1 \qquad \dfrac{\text{Maria}}{N : m}\ Lx}{\begin{array}{c} S : \ \lambda x[A(x)(p)](m) \\ =_{red.\beta}\ A(m)(p) \end{array}}\ /E$$

Figura 6: Derivação de “Pedro ama Maria” equivalente à com R1 e R3

Para se chegar a um diagrama equivalente ao da derivação com R1, R2 e R4, na Figura 3, é necessário não apenas a suposição de uma variável que consome um dos argumentos de “ama”, mas é preciso supor também uma variável que se aplique a “Pedro”

para que, depois de 'descarregada' a suposição, tenhamos transformado "Pedro" numa função que toma um '$N\backslash S$' para resultar num 'S', como podemos ver na Figura 7.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\text{Pedro}}{N : p}\,Lx \quad [N\backslash S : P]^1}{S : P(p)}\,\backslash E \quad \dfrac{\dfrac{\text{ama}}{(N\backslash S)/N : A}\,Lx \quad [N : x]^2}{N\backslash S : A(x)}\,/E \quad \dfrac{\text{Maria}}{N : m}\,Lx}{S/(N\backslash S) : \lambda P[P(p)]}\,/I^1}{S : \begin{array}{l}\lambda P[P(p)](A(x)) \\ =_{red.\beta}\ A(x)(p)\end{array}}\,/E}{S/N : \lambda x[A(x)(p)]}\,/I^2}{S : \begin{array}{l}\lambda x[A(x)(p)](m) \\ =_{red.\beta}\ A(m)(p)\end{array}}\,/E$$

Figura 7: Derivação de "Pedro ama Maria" equivalente à com R1, R2 e R4

3. Comparações

Para um lingüista acostumado com as árvores de uma Gramática de Estrutura Sintagmática, não é difícil ver, num diagrama como o da Figura 1, os itens lexicais claramente representados na parte superior da derivação. Logo abaixo desses itens lexicais, também são claras as representações das operações sintáticas e semânticas reunindo esses itens lexicais em expressões cada vez mais complexas. Como já foi dito, mesmo em diagramas como os da Figura 2 e 3, com derivações incomuns para uma Gramática de Estrutura Sintagmática, podemos relacionar as regras de permutação (R3) e de promoção (R4) a uma operação que afeta a estrutura prosódica da expressão, trocando de lugar fronteiras fonológicas e reagrupando seus constituintes entoacionais.

Já em relação aos diagramas das Figuras 6 e 7, a que tipo de coisas poderíamos relacionar as suposições entre colchetes?

Numa primeira tentativa de responder esta pergunta, poderíamos querer ver nestas suposições as chamadas categorias vazias da Gramática Gerativa, já que elas parecem introduzir apenas uma entidade da representação semântica (uma variável que será posteriormente abstraída pelo operador-λ) que não corresponde a nenhuma expressão lingüística. No entanto, na Gramática Gerativa, as categorias vazias são itens lexicais. Nas derivações da Gramática Categorial, os itens lexicais são claramente representados, no topo do diagrama, como uma expressão lingüística sobre uma barra horizontal, debaixo da qual constam sua categoria e sua representação semântica. É apenas na introdução dos itens lexicais que aparecem expressões lingüísticas sobre estas barras; no restante da derivação sobre as barras ficam apenas as premissas das conclusões localizadas sob a barra. Assim, era de se esperar que, se as suposições fossem categorias vazias, elas fossem introduzidas nos diagramas através de uma barra com o rótulo lexical com a parte acima da barra vazia; portanto, fica impossível relacionar estas suposições às categorias vazias.

Além disso, ainda que essa constatação não seja feita em nenhum lugar, os itens lexicais parecem corresponder a axiomas. Ou seja, eles precisam ser definidos em algum lugar, e eles não podem ser derivados a partir de nenhum outro teorema da Gramática Categorial. Mas, nas regras de introdução dos conectivos, as suposições não correspondem a nenhum axioma ou teorema; pelo contrário, nos diagramas, a suposição nunca aparece

debaixo de uma barra horizontal, como observamos. Dessa maneira, não é apenas às categorias vazias que as suposições não podem corresponder: elas também não podem ser comparadas a nenhum item lexical.

Contudo, mesmo que tivéssemos descoberto uma natureza mais lingüisticamente motivada para o papel das suposições nas regras de introdução de conectivos, ainda precisaríamos encontrar a motivação lingüística que justificasse a maior quantidade de regras empregadas nos diagramas do cálculo de Lambek, em relação aos diagramas equivalentes da versão das regras de redução.

Se observarmos os diagramas das Figuras 2 e 3, podemos constatar que ambos são formados apenas através de três aplicações de regras: uma aplicação de R3 e duas de R1, no diagrama da Figura 2, e uma aplicação de R4, uma de R2 e uma de R1, no diagrama da Figura 3. E em ambos os casos, é fácil relacionar as aplicações das regras unárias (R3 e R4) a operações fonológicas que afetam a organização dos constituintes prosódicos: em ambas as derivações a aplicação das regras unárias faz com que a estrutura prosódica se torne '((Pedro ama) Maria)'; ao contrário da derivação na qual essas regras não atuam, na Figura 1, cuja estrutura prosódica é '(Pedro (ama Maria))'.

Já nas derivações equivalentes do cálculo de Lambek, nas Figuras 6 e 7, as mesmas derivações são realizadas respectivamente através de quatro e seis aplicações das regras. Aqui, fica impossível equiparar cada uma dessas aplicações das regras a qualquer operação fonológica, já que há uma mesma diferença prosódica a ser relacionada à aplicação de uma eliminação e de uma introdução, no diagrama da Figura 6, e à aplicação de duas eliminações e duas introduções, no diagrama da Figura 7.

A diferença fica ainda mais ressaltada quando a derivação é apresentada com o cálculo de seqüentes, de Gentzen, como se pode ver na Figura 8.[9]

$$\dfrac{\dfrac{}{N : m \Rightarrow N : m}\text{Id} \qquad \dfrac{\dfrac{}{N : p \Rightarrow N : p}\text{Id} \qquad \dfrac{}{S : A(m)(p) \Rightarrow S : A(m)(p)}\text{Id}}{N : p, N\backslash S : A(m) \Rightarrow S : A(m)(p)}\backslash\text{Esq}}{N : p, (N\backslash S)/N : A, N : m \Rightarrow S : A(m)(p)}/\text{Esq}$$

Figura 8: Derivação de "Pedro ama Maria" com seqüentes

Na derivação da Figura 8, fica difícil para um lingüista reconhecer o que poderia corresponder a um item lexical, e principalmente identificar a entrada lexical relativa ao verbo "ama": como a introdução de "Maria" e de "Pedro" é feita por duas instâncias do axioma da identidade (Id), não seria de esperar que "ama" também fosse introduzido por outra instância do mesmo axioma? No cálculo de seqüentes, não. Apenas as expressões atômicas aparecem nas instâncias do axioma de identidade: os nomes "Maria" e "Pedro", e a sentença "Pedro ama Maria". Como a categoria de "ama" é funcional, ele só aparece na conclusão da última inferência. Mas observe que, nessa representação, as expressões lingüísticas propriamente ditas sequer aparecem nos diagramas: vemos apenas suas categorias e suas representações semânticas, mas em lugar nenhum podemos perceber as expressões "Pedro", "ama" ou "Maria".

E se já é difícil identificar os próprios itens lexicais, é ainda mais difícil relacionar com o diagrama de seqüentes as operações de concatenação prosódica apontadas anteriormente. A maneira mais simples de apresentar as derivações equivalentes às com a permutação (R3) e com promoção e composição (R4 e R2), seria demonstrando-as separadamente, e depois substituindo na derivação os itens lexicais pelas equivalências

[9] Infelizmente, por falta de espaço, não será possível apresentar aqui a formalização do cálculo de Lambek com seqüentes, que pode ser encontrada em [Moortgat, 1988], [Morrill, 1994] e [Carpenter, 1997].

demonstradas, através da regra de corte (*cut rule*). Mas a que tipo de operação lingüística poderiam corresponder essa demonstração paralela e a própria regra de corte? Com efeito, não é possível relacioná-las lingüisticamente a nada.[10]

4. Conclusão

Através da observação de como os itens lexicais e uma operação lingüística (a concatenação de constituintes prosódicos) poderiam ser identificados nos diagramas de derivação da Gramática Categorial, o que se conclui é que não apenas "as derivações no cálculo associativo de Lambek são representadas mais economicamente por derivações da dedução natural ao estilo de Prawitz" [Morrill, 1994, p. 80], mas que sua versão com os seis pares de regras de redução oferecem um ambiente mais propício para a reflexão mais lingüisticamente motivada.

Assim, mesmo que "a perspectiva dos seqüentes seja uma base particularmente lúcida para a discussão de questões essenciais como as de derivabilidade e de decidibilidade" [Moortgat, 1988, p. 27], a equivalência entre os dois sistemas garante que as descobertas feitas para a versão com os seqüentes possa ser imediatamente transferida para a versão das regras de redução, que é mais adequada ao trabalho do lingüista.

$$\dfrac{\dfrac{}{\text{Pedro} - N : p}\,Lx \quad \dfrac{\dfrac{}{\text{ama} - (N\backslash S)/N : A}\,Lx \quad \dfrac{}{\text{Maria} - N : m}\,Lx}{(\text{ama Maria}) - N\backslash S : A(m)}\,R1}{(\text{Pedro (ama Maria)}) - S : A(m)(p)}\,R1$$

Figura 9: Explicitando a concatenação prosódica apenas com R1

$$\dfrac{\dfrac{}{\text{Pedro} - N : p}\,Lx \quad \dfrac{\dfrac{}{\text{ama} - (N\backslash S)/N : A}\,Lx}{\text{ama} - N\backslash(S/N) : \lambda x[\lambda y[A(y)(x)]]}\,R3 \quad \dfrac{}{\text{Maria} - N : m}\,Lx}{(\text{Pedro ama}) - S/N : \begin{array}{l}\lambda x[\lambda y[A(y)(x)]](p)\\ =_{red.\beta}\ \lambda y[A(y)(p)]\end{array}}\,R1$$
$$\dfrac{}{((\text{Pedro ama}) \text{ Maria}) - S : \begin{array}{l}\lambda y[A(y)(p)](m)\\ =_{red.\beta}\ A(m)(p)\end{array}}\,R1$$

Figura 10: Explicitando a concatenação prosódica com R1 e R3

$$\dfrac{\dfrac{\dfrac{}{\text{Pedro} - N : p}\,Lx \quad \dfrac{}{\text{ama} - (N\backslash S)/N : A}\,Lx \quad \dfrac{}{\text{Maria} - N : m}\,Lx}{\text{Pedro} - S/(N\backslash S) : \lambda P[P(p)]}\,R4}{(\text{Pedro ama}) - S/N : \begin{array}{l}\lambda x[\lambda P[P(p)](A(x))]\\ =_{red.\beta}\ \lambda x[A(x)(p)]\end{array}}\,R2$$
$$\dfrac{}{((\text{Pedro ama}) \text{ Maria}) - S : \begin{array}{l}\lambda x[A(x)(p)](m)\\ =_{red.\beta}\ A(m)(p)\end{array}}\,R1$$

Figura 11: Explicitando a concatenação prosódica com R1, R2 e R4

[10] Novamente por limitação de espaço, também não apresentaremos os diagramas de seqüentes equivalentes às derivações das Figuras 2 e 3. Um diagrama com exemplo do uso da regra de corte para introdução da promoção do sujeito pode ser encontrado em [Carpenter, 1997, p. 147].

Nesse sentido, para encerrar, vamos apresentar uma adpatação para o estilo de Prawitz da representação que [Morrill, 1994, pp. 110–129] desenvolve usando o estilo de Ficht. As principais diferenças desta representação para as apresentadas antes são duas: 1) os itens lexicais ocupam o lugar das premissas nas demonstrações, e 2) as expressões lingüísticas aparecem explicitamente concatenadas. Os diagramas correspondentes às derivações da Figura 1, 2 e 3 podem ser vistos, respectivamente, nas Figuras 9, 10 e 11.

Nestes três últimos diagramas, podemos ver claramente os dois principais pontos ressaltados durante as discussões apresentadas aqui:

- Os itens lexicais correspondem a axiomas, introduzidos sob uma barra sem nada sobre ela; do ponto de vista lingüístico, a conseqüência é que os itens lexicais são independentes: eles não dependem diretamente de nenhuma operação da Gramática Categorial — pelo contrário, são os itens lexicais que afetam a análise lingüística representada nas derivações, assim como o lexicalismo preconiza.
- As operações de concatenação das expressões lingüísticas estão explicitamente expressas, assim como as operações de combinação categorial e de unificação das representações semânticas; assim, em cada passo da derivação, sabemos de cada expressão lingüística construída sua categoria e sua interpretação semântica, relacionando claramente expressões lingüísticas e suas respectivas interpretações semânticas, exatamente como exige a composicionalidade.

Dessa maneira, ainda que aparentemente a discussão acima tenha sido fundamentada pela ontologia das entidades lingüísticas (mais especificamente, pela natureza axiomática dos itens lexicais) e pela epistemologia das operações lingüísticas (não apenas as de combinação categorial e de interpretação semântica, mas também as de aglutinação de constituintes prosódicos), as conclusões a que acabamos de chegar afetam diretamente a representação desse tipo de conhecimento na elaboração de analisadores gramaticais que sirvam como modelo para o comportamento lingüístico humano: a implementação de analisadores para Gramáticas Categoriais que se pretendam psicologicamente realísticos deve representar os itens lexicais como axiomas de uma álgebra livre.

Referências

Ajdukiewicz, K. (1935). Die syntaktische konnexität. *Studia Philosophica*, 1:1–27.

Bar-Hillel, Y. (1953). A quasi-arithmetical notation for syntactic description. *Language*, 29:47–58.

Carpenter, B. (1997). *Type-Logical Semantics*. The MIT Press, Cambridge, Massachusetts.

Cohen, J. M. (1967). The equivalence of two concepts of categorial grammar. *Information and Control*, 10:475–484.

Moortgat, M. (1988). *Categorial Investigations — Logical and Linguistic Aspects of the Lambek Calculus*. Foris, Dordrecht.

Morrill, G. V. (1994). *Type Logical Grammar — Categorial Logic of Signs*. Kluwer, Dordrecht.

Neto, J. B. (1999). Introdução à gramática categorial. UFPR, Curitiba.

Steedman, M. (1988). Combinators and grammars. In Oehrle, R., Bach, E., and Wheeler, D., editors, *Categorial Grammars and Natural Language Structures*, pages 417–442. Reidel, Dordrecht.

Wood, M. M. (1993). *Categorial Grammars*. Routledge, London.

Modelos de Linguagem N-grama para Reconhecimento de Voz com Grande Vocabulário

Ênio Silva, Marcus Pantoja, Jackline Celidônio e Aldebaro Klautau

[1]Laboratório de Processamento de Sinais – Universidade Federal do Pará
DEEC-CT, Belém, PA, 66075-900, Brasil
http://www.laps.ufpa.br
E-mail: a.klautau@ieee.org

Abstract. *This work describes preliminary results on N-gram language models applied to Brazilian Portuguese. The project is part of an effort to develop a large vocabulary continuous speech recognition system, where language modeling plays a fundamental role. We present a brief summary of state-of-art techniques, including the recently proposed interpolated additive (AI) model. We also describe simulation results, which show that the AI model is competitive with some well-established techniques.*

Resumo. *Este trabalho apresenta resultados preliminares acerca do uso de modelos estatísticos N-grama para o português brasileiro. O mesmo se insere no âmbito do desenvolvimento de um sistema de reconhecimento de voz com suporte a grandes vocabulários, onde a modelagem da linguagem é um aspecto fundamental. Apresentamos um breve sumário das técnicas do estado-da-arte, dentre as quais o modelo aditivo interpolado, recentemente proposto. Descrevemos também, de forma comparativa, os resultados obtidos por essas técnicas.*

1. Introdução

A modelagem da linguagem é ingrediente essencial de muitos sistemas computacionais, tais como reconhecimento de voz. Geralmente, os sistemas de reconhecimento de voz (SRV) são baseados em cadeias escondidas de Markov (HMMs, de *hidden Markov models*) [Huang et al., 2001]. Esses sistemas convertem o sinal de voz digitalizado em uma matriz $\mathbf{X}$ de *parâmetros*, e buscam a seqüência de palavras W que maximiza a probabilidade condicional

$$\hat{W} = \arg\max_{W} p(W|\mathbf{X}).$$

Na prática, usa-se a regra de Bayes para implementar a busca através de:

$$\hat{W} = \arg\max_{W} p(W|\mathbf{X}) = \arg\max_{W} \frac{p(\mathbf{X}|W)p(W)}{p(\mathbf{X})} = \arg\max_{W} p(\mathbf{X}|W)p(W),$$

com $P(\mathbf{X})$ sendo desprezado pois não depende de W. Para cada W, os valores de $P(\mathbf{X}|W)$ e $P(W)$ são fornecidos pelos *modelos acústico* e *de linguagem* (ou *língua* [Pessoa et al., 1999b]), respectivamente. Ambos modelos são imprescindíveis em SRV, mas esse trabalho concentra-se nos modelos de linguagem.

Modelos estatísticos de linguagem fornecem a probabilidade de uma seqüência de palavras $W = w_0 \dots w_l$, a qual também será representada por w_0^l e chamada genericamente de *sentença*. Nós assumimos que w_0 é um símbolo para o início da sentença consistindo de $l-1$ palavras, e w_l é um símbolo para o final da sentença. O modelo de linguagem mais utilizado para aplicações em reconhecimento de voz usa a aproximação *n*-grama , a qual assume que a distribuição de probabilidade para a palavra atual depende somente das $n-1$ palavras precedentes:

$$p(w_1^l|w_0) = \prod_{i=1}^{l} p(w_i|w_0^{i-1}) \approx \prod_{i=1}^{l} p(w_i|w_{i-n+1}^{i-1}).$$

Ressalta-se que a probabilidade para o símbolo final da sentença será avaliada no fim da sentença como se fosse uma outra palavra, enquanto que o começo da sentença é tratado apenas como uma informação do histórico (ou *contexto*).

Na criação do modelo de linguagem é desejável então encontrar estimativas ótimas para probabilidades condicionadas a cada contexto. A principal dificuldade em encontrar essas estimativas provém da esparsidade dos dados do treinamento. Uma vez que muitas palavras são nunca ou raramente observadas, suas estimativas não são confiáveis. Para um reconhecedor de voz, palavras que possuem probabilidade zero nunca serão reconhecidas nem que elas sejam acusticamente plausíveis. Isso é chamado de *problema da freqüência zero*. Existem muitas técnicas de suavização que buscam assegurar que todas as palavras, mesmo as que não apareçam no conjunto de treino, possuam probabilidade positiva.

Para melhor estabelecer os objetivos do presente trabalho, alguns conceitos importantes são descritos a seguir. Um texto T é uma coleção de sentenças e sua probabilidade $p(T)$ é o produto da probabilidade de sentenças individuais (assume-se independência estatística entre as sentenças). Para avaliar a qualidade de um modelo de linguagem em T, pode-se usar a entropia cruzada (também chamada *per-word coding length* ou *cross-entropy*)

$$H_p(T) \stackrel{\text{def}}{=} \frac{1}{W_T} \log_2 \left(\frac{1}{p(T)} \right),$$

onde W_T denota o número de palavras em T. Note que se uma probabilidade zero é atribuída a uma palavra que aparece no texto, $H_p(T)$ é infinita. A partir de $H_p(T)$, pode-se definir a *perplexidade* como

$$\text{PP} \stackrel{\text{def}}{=} 2^{H_p(T)}.$$

A perplexidade pode ser entendida como o número médio de diferentes (e equiprováveis) palavras que podem seguir uma dada palavra, de acordo com o modelo de linguagem adotado. Por exemplo, $\text{PP} = 10$ em um SRV para dez dígitos (0 a 9). Para SRV da língua inglesa, com vocabulários de tamanho superior a 20.000 palavras, PP costuma variar entre 100 e 250. Para uma dada tarefa de reconhecimento de voz, objetiva-se encontrar modelos de linguagem que conduzam a baixas perplexidades e custo computacional reduzido.

Considerando-se o SRV como um todo, a medida mais comum de avaliação é a taxa de palavras erradas (WER, de *word error rate*). Pode-se avaliar modelos de linguagem mantendo-se o modelo acústico fixo, e observando-se como as diferentes técnicas impactam a WER. Contudo, essa estratégia possui um custo computacional alto, sendo

comum a utilização da perplexidade nos estágios iniciais do desenvolvimento de modelos de linguagem para SRV. Isso é justificado pela forte correlação entre WER e PP ou, equivalentemente, $H_p(T)$, como indica a expressão[1]

$$\text{WER} \approx -12.37 + 6.48 \log_2(\text{PP}) = -12.37 + 6.48 H_p(T).$$

Assim, o principal objetivo desse trabalho é a obtenção de bons modelos de linguagem para o português brasileiro, e a avaliação será feita através do decréscimo de PP ou $H_p(T)$.

Ressalta-se que há diversos grupos de pesquisa desenvolvendo SRV em universidades como UFSC [Seara et al, 2003], PUC-RJ [Santos and Alcaim, 2002], INATEL [Ynoguti and Violaro, 1999], e PUC-RS [Fagundes and Sanches, 2003], mas há relativamente poucos trabalhos publicados acerca de modelos de linguagem para SRV usando o português brasileiro [Pessoa et al., 1999a, Pessoa et al., 1999b].

Este artigo encontra-se organizado da seguinte forma. Na Seção 2 faz-se uma breve revisão dos mais importantes modelos de linguagem adotados em reconhecimento de voz. Essa revisão é fortemente baseada no trabalho de nossos colaboradores [Jevtic and Orlitsky, 2003]. Na Seção 3 são apresentados resultados de simulação para algumas das técnicas discutidas, comparando-as de acordo com a abordagem adotada em [Chen and Goodman, 1999]. Na Seção 4 são apresentadas as conclusões do trabalho.

2. Estimação dos Modelos de Linguagem N-grama

Entre as primeiras aproximações para o *problema da freqüência zero* encontra-se a *suavização aditiva*, que remonta da época de Laplace [de Laplace, 1816]. Dado um conjunto de símbolos V, denotamos $c(v)$ o número de vezes que o símbolo $v \in V$ foi gerado. Esses estimadores atribuem para cada símbolo $w \in V$ a probabilidade

$$p_{add}(w) \stackrel{\text{def}}{=} \frac{c(w) + \delta}{\sum_{v \in V} (c(v) + \delta)}.$$

Essa equação é conhecida como lei de sucessão de Laplace (veja, e.g., [Jeffreys, 1939, Witten and Bell, 1991]). A regra "add-one" usa a lei de Laplace de sucessão com $\delta = 1$ para estimar a probabilidade da próxima palavra. Este foi um dos primeiros métodos empregados na modelagem da linguagem, mas em [Gale and Church, 1994], os autores mostraram experimentalmente que a mesma tem uma baixa performance. Um método que supera a regra "add-one"na tarefa de modelagem da linguagem é a regra "add-small-delta". Nesse caso, usa-se um subconjunto dos dados de treino (chamada *validação*) para encontrar o δ que maximiza a probabilidade desse subconjunto.

Em geral, as regras "add-small-delta"e "add-one"são eficientes quando todas as probabilidades são diferentes de zero, o que não é o caso em modelagem da linguagem para reconhecimento de voz. Para driblar a esparsidade dos dados, a maioria dos modelos populares de linguagem usa o conceito de *back-off*. Ao invés de remanejar a probabilidade, a distribuição do contexto mais amplo é usada, pois os mesmos possuem estimativas

[1]Obtida por W. Fisher a partir do estudo de diversos SRV, e divulgada em reunião organizada pelo NIST / EUA em 2000. Veja *http://www.isip.msstate.edu/publications/courses/ece_8463/*.

mais robustas. Por exemplo, de um contexto com as $n-1$ palavras mais recentes, recorre-se a um contexto com as $n-2$ palavras mais recentes. A recursão poderia finalizar em uma distribuição uniforme.

Chen e Goodman [Chen and Goodman, 1999] distinguem duas implementações de back-off, a *estrita* e a *interpolada*, e concluem que a interpolada leva a melhores resultados do que a estrita. Assim, neste artigo considera-se somente a variante interpolada:

$$p(w_i|w_{i-n+1}^{i-1}) = \overline{\lambda} \cdot p_0(w_i|w_{i-n+1}^{i-1}) + \lambda \cdot p(w_i|w_{i-n+2}^{i-1}),$$

onde λ é o parâmetro de interpolação e p_0 a estimativa inicial para a probabilidade desejada. Há vários métodos para balanceamento da distribuição do contexto total e seu back-off. A seguir, nós apresentamos alguns dos mais populares. Mais detalhes podem ser encontrados em [Chen and Goodman, 1999].

2.1. Modelo de Jelinek-Mercer

Jelinek e Mercer [Jelinek and Mercer, 1980] descreveram uma classe geral de modelos N-grama que interpolam diferentes cadeias de Markov:

$$p(w_i|w_{i-n+1}^{i-1}) = \sum_{j=0}^{n-1} \lambda_j \cdot p_{ML}(w_i|w_{i-j}^{i-1}),$$

onde $\sum_{j=0}^{n-1} \lambda_j = 1$, $p_{ML}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i})}{c(w_{i-n+1}^{i-1})}$ e $c(w_{i-n+1}^{i})$ representa quantas vezes w_{i-n+1}^{i} foi observada durante o treino. Os dados para treinamento são divididos em dois subconjuntos disjuntos: *treino* e *validação*. A estimativa de máxima verossimilhança (MLE, de *maximum likelihood estimation*) dos dados do conjunto *treino* é usada para obter as probabilidades p_{ML} de cada nível, e os parâmetros λ de interpolação são otimizados para maximizar a probabilidade do conjunto *validação*. A performance de suavização de Jelinek-Mercer é relativamente fraca quando se usa o mesmo λ para todos os contextos, mas inviável caso se adote um λ diferente para cada contexto. Uma solução de compromisso particularmente útil é a interpolação de forma *hierárquica* [Brown et al., 1992]:

$$\begin{aligned} p_{interp}(w_i|w_{i-n+1}^{i-1}) = \quad & \lambda_{w_{i-n+1}^{i-1}} p_{ML}(w_i|w_{i-n+1}^{i-1}) + \\ & (1 - \lambda_{w_{i-n+1}^{i-1}}) p_{interp}(w_i|w_{i-n+2}^{i-1}). \end{aligned}$$

A definição hierárquica permite agrupar λ's para vários contextos similares, separadamente a cada nível. Após isso, a mesma estima conjuntamente os valores ótimos através do algoritmo "expectation-maximization" (EM). O critério original usado para agrupamento em [Brown et al., 1992] foi o número de vezes que o contexto foi observado (contagem total). Assumia-se que um contexto que ocorre um grande número de vezes conduz a uma estimativa mais confiável. O parâmetro crítico que deve ser escolhido é o número dos contextos que serão agrupados, ou o número de parâmetros de interpolação livres que devem ser estimados. Ressalta-se contudo, que estes números dependem do tamanho dos dados do treinamento. Chen mostra em sua tese [Chen, 1996] que é melhor usar a contagem média da palavra como um critério para se aglomerar. Este critério é também muito menos sensível à mudança do tamanho do conjunto, comparado ao critério da contagem total.

2.2. Desconto Linear e Absoluto

Ney, Essen e Kneser [Ney and Essen, 1991, Ney et al., 1994] discutiram que todas as palavras em contextos mais longos são superamostradas ("oversampled") e que há duas maneiras gerais de descontar (ou de reduzir suas probabilidades) para compartilhar a probabilidade com as palavras não observadas do back-off: *desconto linear* e *absoluto*. No desconto linear a MLE dos contextos são descontadas proporcionalmente às probabilidades (escaladas) e a probabilidade descontada total é dada ao back-off (isso corresponde à suavização de Jelinek-Mercer com um único conjunto).

No desconto absoluto, todas as palavras são descontadas por uma constante aditiva igual:

$$p_{abs}(w_i|w_{i-n+1}^{i-1}) = \begin{cases} \frac{c(w_{i-n+1}^{i})-D}{c(w_{i-n+1}^{i-1})}, & \text{se } c(w_{i-n+1}^{i}) > 0 \\ \frac{D \cdot N_{1+}(w_{i-n+1}^{i-1}\cdot)}{c(w_{i-n+1}^{i-1})} p_{abs}(w_i|w_{i-n+2}^{i-1}), & \text{senão} \end{cases} \quad (1)$$

Assume-se que $0 < D < 1$ e $N_{1+}(w_{i-n+1}^{i-1}\cdot)$ é o número de palavras diferentes que foram observadas uma ou mais vezes seguindo w_{i-n+1}^{i-1}. Os autores mostrarm que o desconto absoluto (Equação 1) tem um desempenho melhor do que o desconto linear. Entretanto, quando agrupamentos de contextos são usados para o desconto linear, o desempenho é semelhante.

2.3. Modelo Kneser-Ney

Kneser e Ney [Kneser and Ney, 1995] aperfeiçoaram o modelo de desconto absoluto, impondo uma restrição à distribuição do back-off, forçando distribuições de ordem mais alta a terem as mesmas marginais dos dados de treinamento

$$\sum_{w_{i-n+1}} p_{KN}(w_{i-n+1}^{i}) = \frac{c(w_{i-n+2}^{i})}{N}.$$

De acordo com o modelo, a distribuição back-off é proporcional não ao número de vezes que a palavra foi observada no contexto, mas sim ao número de diferentes contextos nos quais foi observada

$$p_{KN}(w_i|w_{i-n+2}^{i-1}) = \frac{N_{1+}(\cdot w_{i-n+2}^{i})}{N_{1+}(\cdot w_{i-n+2}^{i-1}\cdot)}.$$

Isto produziu uma grande melhoria de desempenho.

2.4. Variação do Modelo de Kneser-Ney

Ney *et al.* [Ney et al., 1997] sugeriram uma variação do desconto absoluto que basicamente usa dois descontos: D_1 para os símbolos observados uma vez, e D_{2+} para aqueles observados duas ou mais vezes. Chen e Goodman em [Chen and Goodman, 1999] mostraram que três constantes D_1, D_2 e D_{3+} tem desempenho consistentemente superior. Ressaltamos então que o modelo de linguagem de Kneser-Ney com três parâmetros de descontos é muitas vezes considerado o melhor algoritmo para estimar um modelo N-grama.

2.5. Modelo Aditivo Interpolado

Em [Jevtic and Orlitsky, 2003], foi proposto o modelo *aditivo interpolado* (AI). Para cada contexto w_{i-n+1}^{i-1} de palavras observadas nos dados de treino, usa-se uma constante aditiva $\delta \in (-1, +\infty)$ para suavizar a distribuição.

$$p(w_i|w_{i-n+1}^{i-1}) = \begin{cases} \frac{c(w_{i-n+1}^{i})+\delta}{c(w_{i-n+1}^{i-1})+N_{1+}(w_{i-n+1}^{i-1}\cdot)\delta}, & c(w_{i-n+1}^{i}) > 0 \\ 0, & c(w_{i-n+1}^{i}) = 0. \end{cases}$$

Já para as palavras não observadas nos dados de treino, usa-se a aproximação interpolada de Jelinek-Mercer:

$$\begin{aligned} p_{opt}(w_i|w_{i-n+1}^{i-1}) = & \quad \lambda_{w_{i-n+1}^{i-1}} p(w_i|w_{i-n+1}^{i-1}) \\ & +(1-\lambda_{w_{i-n+1}^{i-1}}) p_{opt}(w_i|w_{i-n+2}^{i-1}). \end{aligned}$$

O final da recursão é uma distribuição uniforme. Esta formulação permite usar o algoritmo EM para estimar λ's em níveis diferentes, da mesma maneira usada no modelo de Jelinek-Mercer.

3. Resultados

Nesta seção são apresentados os resultados de simulações. Foi utilizado um corpus[2] do português brasileiro constituído majoritariamente por textos de um jornal e formatado usando XML. O corpus tem aproximadamente 30 milhões de linhas, das quais foram retirados os tags XML. A pontuação foi substituída por tags especiais, tais como <EXCLAMAÇÃO>, <VÍRGULA>, etc. Os dados foram separados em três conjuntos disjuntos para treino, validação e teste. Tanto o conjunto de *validação* quanto o de teste foram mantidos em 2500 sentenças.

Na Figura 1 encontram-se os resultados de um experimento preliminar usando bigramas e trigramas estimadas através do método "default" do software HTK (*http://htk.eng.cam.ac.uk/*). Esses resultados são compatíveis com os obtidos para a língua inglesa em simulações semelhantes, onde a perplexidade situa-se em torno de 100 a 250.

No intuito de aperfeiçoar os modelos obtidos com o HTK, lançamos mão do software desenvolvido por Nikola Jevtic (também usado em [Jevtic and Orlitsky, 2003]). De forma similar à metodologia em [Chen and Goodman, 1999], comparamos as técnicas mais populares em função do aumento no tamanho da seqüência de treino. Para os métodos que requerem "clustering" dos parâmetros de interpolação (Jelinek-Mercer e AI), os modelos foram construídos para diversos tamanhos de cluster e foi escolhido o que melhor se adapta a um segundo conjunto *validação* (também de 2500 sentenças).

Os resultados das novas simulações com trigramas são mostrados na Figura 2. Seguindo o formato adotado em [Chen and Goodman, 1999, Jevtic and Orlitsky, 2003], todos os gráficos mostram a diferença relativa na entropia $H_p(T)$ quando o método é comparado com o resultado mostrado Figura 1 (obtido com o HTK). Pode-se observar

[2]Gentilmente fornecido pelo Professor Ticiano Monteiro do CESUPA-PA.

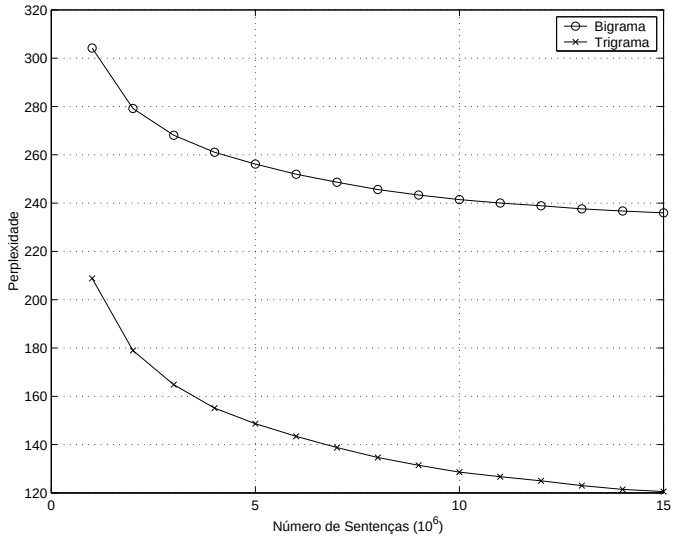


Figura 1: Evolução da perplexidade PP com o aumento dos dados para treino, para bigramas e trigramas.

uma melhoria de desempenho, com exceção do método de desconto absoluto. Verifica-se também que o método AI apresenta desempenho bem próximo ao do Kneser-Ney modificado. Ressalta-se contudo, que o AI apresenta melhor escalabilidade quando se aumenta a duração do contexto (ou seja, usa-se n-gramas com maior n), de acordo com [Jevtic and Orlitsky, 2003].

4. Conclusões

Este trabalho apresentou um breve sumário das técnicas do estado-da-arte em modelagem de linguagem, dentre as quais o modelo AI, recentemente proposto. Apresentou-se também, de forma comparativa, os resultados obtidos por várias das técnicas mais importantes quando aplicadas a um corpus de português brasileiro. Foi constatado que o modelo AI atinge bons resultados, com um custo computacional relativamente baixo quando comparado a métodos de desempenho similar.

Como todo sistema *data-driven*, o reconhecimento de voz se beneficia da disponibilidade de corpora com grande volume de dados. Existe uma quantidade razoável de textos para estudos de modelagem de linguagem para a língua inglesa, português europeu e outras (vide catálogo do LDC em *http://www.ldc.upenn.edu/*). Todavia, há poucos recursos acessíveis quando se trata do português brasileiro. Essa lacuna é ainda maior quando se trata de voz digitalizada para treinamento do modelo acústico. A inexistência dessas bases de dados não só atrasa as pesquisas em reconhecimento de voz e áreas correlatas, mas também impede que os resultados obtidos por diferentes grupos de pesquisa sejam comparados diretamente.

Futuros desenvolvimentos desse trabalho incluem o aumento da base de dados, melhoria dos algoritmos de estimação de n-gramas e uma ampla comparação entre os algoritmos no tocante à perplexidade, WER e custo computacional.

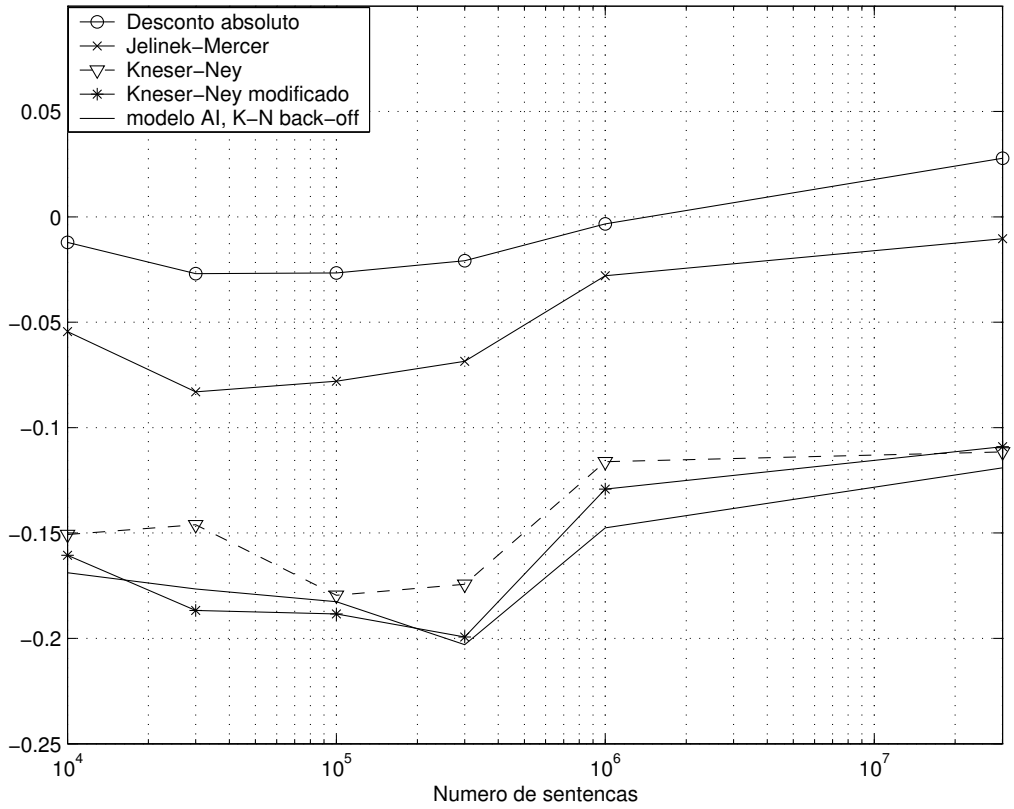


Figura 2: Resultados para trigramas: diferença relativa na entropia $H_p(T)$ em relação ao algoritmo do HTK usado para gerar a Figura 1. Quanto mais negativo o gráfico (menor entropia), melhor o resultado.

Referências

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Lai, J. C., and Mercer, R. L. (1992). An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18:31–40.

Chen, S. F. (1996). *Building Probabilistic Models for Natural Language*. PhD Thesis.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13:359–394.

de Laplace, P. S. (1816). *Essay Philosophique sur la Probabilités*. Courcier Imprimeur, Paris.

Fagundes, R. and Sanches, I. (2003). Uma nova abordagem fonético-fonológica em sistemas de reconhecimento de fala espontânea. *Revista da Sociedade Brasileira de Telecomunicações*, 95.

Gale, W. A. and Church, K. W. (1994). What's wrong with adding one. *Corpus-Based Research Into Language (Oosdijk, N. and de Haan, P., eds)*.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken language processing*. Prentice-Hall.

Jeffreys, H. (1939). *Theory of Probability*. Clarendon, Oxford.

Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of markov source parameters from sparse data. *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397.

Jevtic, N. and Orlitsky, A. (2003). On the relation between additive smoothing and universal coding. *IEEE ASRU*.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1:181–184.

Ney, H. and Essen, U. (1991). On smoothing techniques for bigram-based natural language modeling. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2:825–829.

Ney, H., Essen, U., and Kneser, R. (1994). On structuring probabilistic dependences in stochastic language modeling. *Computer Speech and Language*, 8:1–38.

Ney, H., Martin, S., and Wessel, F. (1997). Statistical language modeling using leaving-one-out. *In Corpus Based Methods in Language and Speech Processing*, pages 174–207.

Pessoa, L., Violaro, F., and Barbosa, P. (1999a). Modelo de língua baseado em gramática gerativa aplicado ao reconhecimento de fala contínua. In *XVII Simpósio Brasileiro de Telecomunicações*, pages 455–458.

Pessoa, L., Violaro, F., and Barbosa, P. (1999b). Modelos da língua baseados em classes de palavras para sistema de reconhecimento de fala contínua. *Revista da Sociedade Brasileira de Telecomunicações*, 14(2):75–84.

Santos, S. and Alcaim, A. (2002). Um sistema de reconhecimento de voz contínua dependente da tarefa em língua portuguesa. *Revista da Sociedade Brasileira de Telecomunicações*, 17(2):135–147.

Seara et al, I. (2003). Geração automática de variantes de léxicos do português brasileiro para sistemas de reconhecimento de fala. In *XX Simpósio Brasileiro de Telecomunicações*, pages v.1. p.1–6.

Witten, I. H. and Bell, T. C. (1991). The zero frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–94.

Ynoguti, C. A. and Violaro, F. (1999). Influência da transcrição fonética no desempenho de sistemas de reconhecimento de fala contínua. In *XVII Simpósio Brasileiro de Telecomunicações*, pages 449–454.

Os tipos de anotações, a codificação, e as interfaces do Projeto Lácio-Web: Quão longe estamos dos padrões internacionais para córpus?

Sandra Maria Aluísio[1,2], Leandro H. M. de Oliveira[1], Gisele Montilha Pinheiro[1]

[1] Núcleo Interinstitucional de Lingüística Computacional (NILC), CP 668, 13560-970 São Carlos, SP, Brasil

[2]ICMC-Universidade de São Paulo, CP 668, 13560-970 São Carlos, SP, Brasil

sandra@icmc.usp.br, {leandroh, gisele}@nilc.icmc.usp.br

Abstract *This paper addresses issues related to the development and standards for public available corpora, including types of annotation, encoding (i.e. forms of representation), tools and database architectures, in connection with the Lacio-Web Project. We also assess whether the decisions made for the Lacio-Web Project conform with the standards and how far are we from a representative Corpus of the Brazilian Portuguese.*

Resumo. *Neste artigo discutimos questões relacionadas aos tipos de anotação, codificação (no sentido de "forma de representação"), ferramentas e arquiteturas para dados, considerados padrões em ambientes de desenvolvimento e disponibilização de córpus. É discutido, também, quão próximas estão as decisões do Projeto Lácio-Web desses padrões e da construção de um Corpus Nacional do Português Brasileiro.*

1. Introdução

Vários córpus foram construídos para a língua inglesa, desde o pioneiro córpus Brown, lançado em 1964 com 1 milhão de ocorrências. Em termos de megacórpus balanceados, tanto o British National Corpus (BNC), para a variante britânica, quanto o American National Corpus (ANC), para a americana, contribuem para o desenvolvimento de ferramentas de processamento de língua natural (PLN) e para a descrição da língua e construção de recursos, tais como dicionários e gramáticas. Além disso, esses córpus impulsionam o desenvolvimento de formatos padrões de anotação e codificação, além de arquiteturas para dados e para ferramentas de manipulação de córpus. São esses padrões internacionais que ajudam a criar grandes córpus que sejam intensivamente usados, reusáveis e extensíveis.

Em [Ide and Brew 2000], a **reusabilidade** (característica de um córpus ser usável em mais de um projeto de pesquisa e por mais de um grupo de pesquisadores) e a **extensibilidade** (isto é, a capacidade de córpus serem melhorados em várias direções, por exemplo, com a provisão de um nível a mais de análise lingüística) são colocadas como dois aspectos a serem considerados em projetos de córpus. Para criar um Córpus Nacional do

Português Brasileiro (CNPB), objetivo de vários pesquisadores no Brasil[1], espera-se que tal megacórpus contemple uma **boa variedade de gêneros, tipos de textos** e **domínios** do conhecimento, inserida numa tipologia textual criteriosa e explícita. Também é desejável que o megacórpus seja **sincrônico** e **contemporâneo** como outros córpus desse tipo, trazendo a produção tanto escrita quanto falada em escala nacional. O córpus deve conter **textos completos** (escritos e transcritos), pois isso viabiliza um tipo especial de estudo, a análise do discurso e do texto. Há, entretanto, uma necessidade que não é de ordem técnica, mas que precede todas as outras, caso esse córpus envolva a disponibilização pública e integral via Web: a obtenção da **autorização de uso dos textos** para pesquisa.

Este artigo apresenta um projeto de desenvolvimento de corpus, o Lácio-Web (LW) [2] [Aluisio et al 2004, Aluisio et al 2003a, Aluisio et al 2003b], em direção à construção de um CNPB. Através do Projeto LW: a) propusemos uma tipologia ortogonal de textos, que privilegia criteriosamente o gênero e o tipo de texto, o domínio e o meio de distribuição; b) obtivemos a autorização de uso dos textos, possibilitando acesso livre desse material via Web; c) criamos uma interface Web de pesquisa e montagem de subcórpus, de modo a atender a maioria dos dados armazenados no cabeçalho das amostras; d) associamos a cada córpus (o LW possui seis tipos diferentes de córpus) um conjunto de ferramentas de processamento lingüístico, muitas das quais já utilizadas em outros projetos do Núcleo Interinstitucional de Lingüística Computacional (NILC)[3]; e e) adequamos o acesso aos córpus, a fim de torná-los de fácil interação entre os usuários especialista e leigos.

Na próxima seção apresentamos o Projeto Lácio-Web, seu status atual e o montante de dados e ferramentas a serem disponibilizados até o final do projeto. Na seção 3, são comentadas as vantagens do uso de XML para criação e manipulação de córpus e de padrões internacionais para codificação e intercâmbio de dados, com vistas à construção de um CNPB. Nessa seção também apresentamos as diferenças e semelhanças desses padrões com as decisões do LW. Na Seção 4, apresentamos as interfaces de pesquisa e de ferramentas.

2. O Projeto Lácio-Web (LW)

LW é um projeto iniciado em 2002, com 30 meses de duração, financiado pelo CNPq, e desenvolvido na Universidade de São Paulo pelo NILC, Instituto de Matemática e Estatística (IME)[4] e Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH)[5]. O Projeto visa ao desenvolvimento de vários tipos de córpus e ferramentas tanto para análise qualitativa (i.e., os dados podem ser utilizados, por exemplo, na construção de dicionários gerais ou terminológicos, ou ainda, na descrição da língua) quanto para a quantitativa (i.e., as estatísticas sobre os dados podem ser utilizadas, por exemplo, na construção de dicionários, etiquetadores morfossintáticos, sintáticos e corretores gramaticais). Os córpus do LW e suas ferramentas (Seção 4) são disponibilizados a partir de uma interface Web. Com respeito aos córpus, o Projeto LW traz: 1) um córpus aberto, sincrônico e contemporâneo de português escrito do Brasil (Lácio-Ref); 2) um córpus fechado, manualmente anotado com etiquetas morfossintáticas (Mac-Morpho); 3) um córpus fechado automaticamente anotado com lemas, etiquetas morfossintáticas e sintáticas para o qual será

[1] Veja em http://www.nilc.icmc.usp.br/iiiencontro/iiiencontro.htm as decisões do III Encontro de Córpus, realizado em 7 de novembro de 2003, no IEL, Unicamp.
[2] http://www.nilc.icmc.usp.br/lacioweb/
[3] http://www.nilc.icmc.usp.br/nilc/index.html
[4] http://www.ime.usp.br/
[5] http://www.fflch.usp.br

usado um *parser* desenvolvido no NILC[6] (Lácio-Sint); 4) um córpus aberto de desvio, contendo textos não revisados segundo os padrões da norma culta (Lácio-Dev); 5) um córpus paralelo aberto contendo textos em inglês e português do Brasil (Par-C), e 6) córpus comparáveis gerados automaticamente a partir de textos do Lácio-Ref e Ref-Ig (um córpus de referência do inglês construído no LW que traz, atualmente, textos originais em inglês do gênero jurídico) (Comp-C). Uma característica que distingue os córpus do LW com outros do português brasileiro é a sua proposta de servirem de *benchmark* para avaliar ferramentas de PLN. É o caso do Mac-Morpho para avaliação de etiquetadores morfossintáticos[7], Comp-C e partes do Lácio-Ref para avaliar métodos automáticos de extração de termos, Par-C para avaliação de alinhadores automáticos e Lácio-Dev para avaliação de corretores gramaticais.

O primeiro lançamento do Projeto se deu em 20/1/2004 e tornou dois córpus disponíveis que são detalhados abaixo: uma versão do Lácio-Ref para pesquisa e geração de subcórpus e o MAC-Morpho para download. O acesso aos córpus se dá após preenchimento de um cadastro.

A versão do Lácio-Ref possui 4.156.816 ocorrências, composta de textos organizados em cinco gêneros (Informativo, Científico, Prosa, Poesia e Drama), vários tipos de textos, vários domínios e alguns meios de distribuição (revista, internet, livro)[8]. O Lácio-Ref é disponibilizado para pesquisa com geração de subcórpus para download e acessado em dois formatos: a) texto com cabeçalho em XML, contendo dados bibliográficos e de classificação textual; e b) texto cru acrescido dos dados relativos ao título e à autoria. Nos subcórpus podem, ainda, ser aplicados três tipos de ferramentas: contadores de freqüência, concordanciadores e etiquetadores.

O MAC-Morpho possui 1.167.183 ocorrências de textos jornalísticos de dez cadernos da Folha de São Paulo, 1994. Essas foram etiquetadas pelo parser Palavras de Eckhard Bick (http://visl.hum.sdu.dk), mapeadas para o conjunto de etiquetas do Projeto Lácio-Web[9] e revisadas manualmente quanto à anotação morfossintática. O MAC-Morpho é disponibilizado para download em 2 formatos: um adequado para pesquisas lingüísticas com o uso de contadores de freqüência ou concordanceadores, por exemplo, e outro adequado ao treinamento de etiquetadores.

Para o lançamento final em junho, que culmina com o fim do suporte financeiro do CNPq, o Lácio-Ref será enriquecido com textos dos gêneros Instrucional, Jurídico, Informativo e Cientifico e contemplará muitos outros tipos de textos, domínios e meios de distribuição, totalizando 8.291.818 ocorrências. Como a tipologia prevê 9 gêneros, os dois restantes (Técnico-Administrativo e De Referência) serão contemplados em projetos de continuação do LW. Também haverá a disponibilização do córpus paralelo Par-C com 646 arquivos de textos em inglês e 646 em português da Revista Pesquisa Fapesp, totalizando 893.283 ocorrências e o lançamento da ferramenta de montagem de córpus comparáveis inglês-português envolvendo o gênero jurídico. Para a construção de córpus comparáveis, foi construído um córpus de referência de textos em inglês (Ref-Ig) do domínio jurídico. Ele

[6] http://www.nilc.icmc.usp.br/nilc/tools/curupira.html
[7] Três etiquetadores morfossintáticos disponíveis na Web foram treinados com o Mac-Morpho, podendo ser utilizados através de uma interface Web no LW. Veja as precisões de cada um em http://www.nilc.icmc.usp.br/lacioweb/ferramentas.htm
[8] Veja os textos do primeiro lançamento, separados por gênero, tipos de texto, domínios e meios de distribuição em http://www.nilc.icmc.usp.br/lacioweb/plancamento.htm
[9] Para saber mais sobre o processo de mapeamento, cf. http://www.nilc.icmc.usp.br/lacioweb/manuais

conta com 15 textos e 22.948 ocorrências e, futuramente, será ampliado. Os córpus Lácio-Dev e Lácio-Sint serão disponibilizados futuramente, como frutos de pesquisas de doutorado e mestrado, respectivamente. No total, o Projeto LW possuirá, no seu segundo lançamento, 5694 arquivos, totalizando 10.375.323 ocorrências.

3. Padrões internacionais para criação e manipulação de córpus

Discutiremos as questões sobre quais os **tipos de anotação**, **codificação** e **ferramentas e arquiteturas para ferramentas e dados**. Os dois primeiros estão bem descritos em [Ide and Romary 2003, Ide et al. 2003] e serão brevemente explicados aqui antes de explorarmos nossas opções. Por sua vez, as ferramentas disponíveis dependem da escolha da representação escolhida e deveriam ser livremente disponíveis e reusáveis para evitar o processo caro de reimplementação de software a cada novo projeto de córpus [Ide and Brew, 2000]; elas também serão exploradas nesse artigo.

Geralmente, distingue-se a anotação de segmentação da anotação lingüística. Na **anotação de segmentação** do texto cru, tem-se: a) *marcação da estrutura geral* – capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos como tabelas e figuras, e b) *marcação da estrutura de subparágrafos* – elementos que são de interesse lingüístico, tais como sentenças, citações, palavras, abreviações, nomes, datas e ênfase. Já na **anotação lingüística** é fornecida a informação lingüística sobre segmentos como etiquetação morfossintática e sintática.

A **representação** se refere ao formato escolhido para explicitar a anotação. É aconselhável que a representação permita a separação entre os dados originais (ou também anotados com a estrutura geral) e anotações para que, por exemplo, possamos aplicar vários tipos de etiquetadores morfossintáticos ou sintáticos num mesmo córpus. Essa estratégia para a arquitetura dos dados que tem sido utilizada em projetos atuais de córpus [Ide and Macleod 2001, Santos and Bick 2000] é diferente da estratégia clássica de adicionar incrementalmente anotações aos dados originais. Mais detalhes dessa discussão estão em [Ide 1998, Ide and Romary 2003].

Um formato bastante usado e que provavelmente será o escolhido para representar a maioria dos córpus é a eXtensible Markup Language (XML) – um padrão internacional para representação e intercâmbio de dados na Web –, pois tem características e extensões úteis para a criação e manipulação de córpus anotados, entre elas: a) XML Links, que permitem endereçar os elementos XML tanto dentro de um mesmo documento como em outros documentos; b) a linguagem XPath e XPoint que, através de predicados, permitem localizar elementos na estrutura de elementos (em árvore) e selecionar fragmentos do texto; c) XSLT, que pode ser usada para converter um documento XML em outro formato; e d) *XML schemas* que estendem o poder dos DTD´s permitindo uma avaliação melhor tanto da forma quanto do conteúdo dos documentos XML [Ide 2000, Ide et al 2000]. XML não é, porém, o único formato para codificar córpus. O IMS Corpus Workbench[10] foi usado no Projeto AC/DC [Santos and Sarmento 2003, Santos and Bick 2000] para disponibilizar córpus do português europeu e brasileiro, e no Projeto Korpus 2000 [Andersen et al 2002], para o dinamarquês.

É interessante contrastar a abordagem gerencial de criação de grandes córpus realizada no projeto AC/DC com uma outra para a criação de um CNPB se decidirmos

[10] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

utilizar um padrão internacional para codificação de córpus como o Corpus Encoding Standard (CES)[11]. O CES é uma aplicação do SGML e possui uma versão mais atual em XML, o XCES. O CES utiliza e adapta os padrões das diretrizes para codificação e intercâmbio de textos eletrônicos TEI[12] para codificar córpus. Na medida em que a abordagem utilizada no projeto AC/DC harmoniza a anotação de segmentação de vários córpus e os centraliza num único site para que os mesmos possam ser pesquisados, a abordagem XCES pretende um desenvolvimento distribuído de anotações e ferramentas de acesso a córpus (uma vez que esse padrão seja adotado por vários projetos de córpus dispersos geograficamente), podendo os dados, anotações e ferramentas de pesquisa serem armazenados em vários servidores dispersos geograficamente. Essa última abordagem favorece a construção de um CNPB, dado que o volume enorme de dados envolvidos e o trabalho de construção de tal recurso inviabilizam que o mesmo seja construído por um único grupo de pesquisa. Além disso, permite utilizar os vários córpus escritos e de fala já construídos. É importante notar, entretanto, que a tarefa de construção de um CNPB não é trivial, envolve altos recursos financeiros e recursos humanos treinados, além de tempo. Essa empreitada, entretanto, faz com que a comunidade de lingüística de córpus envolvida na construção de um CNPB esteja afinada com os padrões de disponibilização de recursos internacionais. O córpus Lácio-Ref e sua tipologia de textos podem fazer parte de um futuro CNPB, entretanto há que se cuidar do balanceamento de gêneros. Foi também uma decisão de projeto privilegiar textos integrais e, assim, o Lácio-Ref não se conforma com as decisões de córpus como o BNC que limita o tamanho das amostras de textos. Tal mudança, contudo, pode ser facilmente realizada.

Várias decisões no projeto do LW ainda estão distantes dos padrões internacionais (como o XCES) tanto com relação à anotação como à codificação. No LW, não há a anotação de grande parte dos elementos da estrutura geral, tais como capítulos, parágrafos, subparágrafos, títulos e notas de rodapé em XML. Porém, eles estão formatados e padronizados para fácil visualização com marcas do tipo quebra de linhas, caixa alta, etc.. Já os elementos gráficos estão todos anotados em XML. No córpus Mac-Morpho, a anotação morfossintática se dá num formato propício para o treinamento de etiquetadores (cada palavra em uma linha juntamente com sua etiqueta); nenhum esforço para separação entre texto cru e anotação foi realizado. Quanto à anotação dos elementos do cabeçalho, esses estão em XML e podem facilmente se adequar às normas do XCES. Um editor de cabeçalho certifica que a geração desse esteja correta e facilita a edição das várias informações. Quanto ao grande trabalho de reescrita de ferramentas de PLN e manipulação de córpus em geral, ele pode ser evitado com a tendência atual de utilizar arquiteturas para construção e manipulação de córpus como GATE[13]. Na abordagem do LW foram colocadas à disposição ferramentas desenvolvidas em projetos realizados durante os 10 anos do NILC, como etiquetadores morfossintáticos, sintáticos, alinhadores automáticos de sentenças e extratores de termos. Assim, privilegia-se o reuso de software.

4. Interfaces de pesquisa e de ferramentas do Portal LW

Uma vez que os principais objetivos do Lácio-Web são a disponibilização de córpus e de ferramentas para análise lingüística e ferramentas de PLN, foram desenvolvidos dois tipos

[11] http://www.cs.vassar.edu/CES/
[12] http://www.tei-c.org/
[13] http://gate.ac.uk

de interfaces: as interfaces de pesquisa de córpus e as interfaces de ferramentas. Apesar de serem interligadas e interdependentes, os objetivos são diferentes. As **interfaces de pesquisa** têm as funções de: i) possibilitar a pesquisa de córpus obedecendo aos critérios de classificação tipológica e bibliográfica dos textos; e ii) com o resultado das pesquisas, promover a montagem de subcórpus. Já as **interfaces de ferramentas** têm como objetivo aplicar as ferramentas disponíveis no LW nos córpus ou subcórpus montados pelos usuários e, conseqüentemente, exibir os resultados. A principal motivação para a criação destas interfaces foi o fato de se garantir o direito dos usuários (especialistas ou não) de pesquisar córpus, bem como o de montar seus subcórpus e sobre eles aplicar, de forma independente, as ferramentas disponíveis.

As interfaces de pesquisa e de ferramentas que discutiremos neste artigo são referentes ao córpus Lácio-Ref e MAC-Morpho, no escopo do LW. Todos os textos pertencentes aos córpus do LW são classificados quanto a dois conjuntos de informações: 1) informações bibliográficas (dados de catalogação: amostragem, título, fonte, status de língua, autoria, tradução, etc.) e 2) informações tipológicas (dados de classificação: gênero e subgênero textual, tipo de texto, domínio e subdomínio, meio de distribuição). Essas informações são únicas e exclusivas de cada texto e são armazenadas num cabeçalho descrito em linguagem XML[14].

A pesquisa dos córpus disponíveis no Portal do LW é depende dessas informações, visto que os campos disponíveis nas interfaces de pesquisa advêm do cabeçalho. Os dados de classificação obedecem a uma hierarquia interna que dita uma relação de subordinação entre os campos (própria da codificação XML), permitindo uma coerência de classificação do texto. Assim como esses, os dados de catalogação também possuem tal relação. Um exemplo: se determinado texto pertence ao domínio Ciências Exatas e da Terra, significa que o subdomínio do texto deverá ser subdomínio subjacente a este domínio, e assim sucessivamente em todas classificações disponíveis no cabeçalho.

Para garantir a execução rápida das consultas dos usuários e a atualização dinâmica da interface, considerando as relações hierárquicas dos campos do cabeçalho, foi necessária a transposição da estrutura do cabeçalho XML numa estrutura de banco de dados relacionais. Essa atualização dinâmica da interface diz respeito à sensibilidade e flexibilidade dos campos de seleção (para pesquisa) pertencentes às interfaces e, conseqüentemente, coerentes com a classificação dos textos. Isso quer dizer que os campos de seleção mostrados na interface são sensíveis ao conteúdo selecionado pelo usuário num dado momento, e podem alterar dinamicamente o conteúdo dos outros campos da pesquisa. A adoção do banco de dados também foi importante para garantir a eficiência e a rapidez no processamento das pesquisas (consultas) no ambiente Web, visto que tais pesquisas exigem a aplicação de vários *joints* (junção de dados relacionados) de tabelas, implementados por meio do uso da linguagem *SQL (Structured Query Language)*. Mencione-se, também, que a utilização de um banco de dados relacional aumenta a segurança e a integridade dos dados armazenados.

---

[14] É importante assinalar alguns pontos aqui pertinentes: a) o Mac-Morpho não tem, até momento, o c-LW inserido nas amostras; b) o tratamento de outros córpus pode gerar a criação de novos dados de catalogação; e c) cada uma das categorias da catalogação possui desdobramentos que se constituem características importantes sobre textos escritos (tipo de autoria, data e local de publicação, link dos dados de tradução para o texto original, etc.).

4.1. As interfaces de pesquisa

Foram definidos três tipos de pesquisas, cujos critérios se nortearam pela expectativa dos tipos de usuários de córpus. De um lado estão os usuários **especialistas,** como lingüistas, gramáticos, analistas do texto e do discurso, sociolingüistas, teóricos da literatura, lingüistas computacionais, lingüistas de córpus, lexicógrafos, terminólogos, cientistas da computação. De outro, os usuários **leigos**: estudantes de toda sorte, revisores de texto, professores de língua, tradutores, historiadores, etc.. À disposição desse público-alvo foram projetadas as seguintes opções de seleção de subcórpus: pesquisa simples, a pesquisa avançada e a pesquisa personalizada.

4.1.1 Pesquisa Simples: é a mais genérica do Portal e, ao mesmo tempo, a que oferece menos opções de seleção aos usuários. O seu caráter genérico se define pela vinculação do sistema de busca com o padrão de nomeação dos arquivos, em que se prevê a seleção de subcórpus pela escolha dos dados relativos à classificação das amostras textuais. Por sua vez, o parâmetro da nomeação de arquivos busca atender às expectativas de um usuário leigo para quem a pesquisa avançada e/ou personalizada podem indicar dificuldade ou não-relevância. Os resultados obtidos (i.e., número de textos recuperados) pela pesquisa simples são, geralmente, extensos.

4.1.2 Pesquisa Avançada: é a intermediária, situada entre a Simples e a Personalizada, permitindo que o usuário refine suas opções e obtenha resultados mais específicos, mas em menor grau que a busca personalizada. Deixa de ser relacionada à nomeação dos arquivos e passa a disponibilizar os dados da catalogação na seleção de subcórpus pelo usuário. Nesse caso, o sujeito que se espera no acesso é o usuário que precisa refinar o seu subcórpus em termos mais definidos de amostras e que é capaz de julgar os textos em termos mais específicos de classificação. Por exemplo, é capaz de dizer que quer apenas textos literários em prosa – biografia, respectivamente o gênero e o subgênero textual. Assim, além dos campos da Pesquisa Simples, os usuários podem selecionar mais dados de classificação (*Supergênero, Gênero* e *Subgênero textual*), bem como os dados de catalogação bibliográficas, como *Nome de Autor*, *Nome do Periódico* e *Caderno*. Os campos *Gênero*, *Subgênero*, *Nome do Periódico* e *Caderno* também possuem conteúdos dinâmicos. A Figura 1 mostra duas telas (A e B) como exemplo deste tipo de seleção.

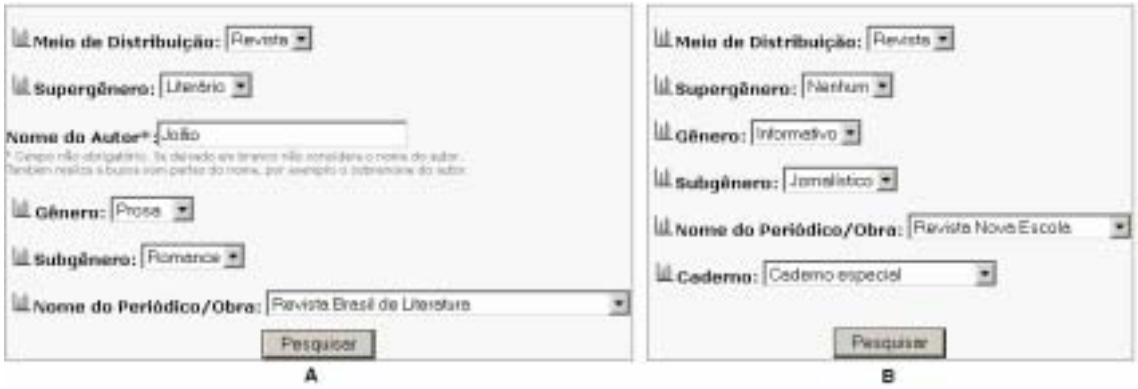


Figura 1 – Telas da Pesquisa Avançada no Portal LW

Observe que nessa ilustração o campo *Caderno* não aparece como opção disponível. Isso acontece porque o *Nome do Periódico* selecionado – a "Revista Brasil de Literatura" – não possui cadernos vinculados. Entretanto, quando o *Supergênero* selecionado é "Literário", o campo *Nome de Autor* é ativado. Em contrapartida, observando a Figura 1-B,

que representa outro exemplo da *Pesquisa Avançada*, verificamos que não aparece o campo *Nome do Autor*; desta vez, o campo *Caderno* está disponível visto que o *Nome do Periódico* "Revista Nova Escola" possui cadernos vinculados.

4.1.3 Pesquisa Personalizada: permite ao usuário refinar sua pesquisa ao máximo, oferecendo opções de seleção que abrigam, em dois grupos, tanto os dados de catalogação como os de classificação. Foi projetada para o usuário especialista, que recorta criteriosamente suas amostras e está a par de todos os detalhes de publicação dos textos que procura. Nessa pesquisa o usuário deve definir detalhadamente o recorte de sua investigação, de maneira que os resultados obtidos sejam de um perfil específico. Novos campos de seleção como: o *Tipo de Amostragem,* o *Tamanho da Amostra*, o *Tipo de Autoria* e o *Tipo Textual* são apresentados ao usuário, sendo que a grande maioria deles possui conteúdos dinâmicos. Como exemplos dessa dinamicidade estão os "novos" campos de *Domínio* e *Subdomínio,* cujos conteúdos são dependentes.

4.2. As interfaces de ferramentas

As interfaces de ferramentas do Portal do LW têm como principal objetivo facilitar a aplicação de ferramentas de análise lingüística aos córpus e/ou subcórpus montados pelos usuários. Sua maior vantagem é a condução do usuário na tarefa de verificar, por meio de ferramentas, a qualidade e relevância dos subcórpus montados. Atualmente, há quatro ferramentas disponíveis no LW. Três delas são aplicadas ao córpus Lácio-Ref e, conseqüentemente, aos subcórpus montados pelos usuários ((a), (b) e (c) abaixo). A outra (um concordanceador) é especificamente aplicada ao córpus etiquetado MAC-Morpho.

a) Contador de Freqüência Padrão: calcula a freqüência com que as palavras ocorrem em um córpus, ferramenta comum em trabalhos com córpus, já que calcular a freqüência de palavras é uma tarefa simples. Porém, o contador de freqüência disponível no Portal LW possui um diferencial relevante, que é o reconhecimento de "lexias complexas dos nomes próprios"[15] e "palavras compostas"[16] para o cálculo das freqüências. Nesse contador, o reconhecimento dos tokens é realizado por um conjunto de regras de formação de palavras ao qual o contador é submetido no momento de sua execução. Além disso, no contador os usuários têm a opção de escolher a ordem de freqüência das palavras (alfabética ou decrescente) e também em qual córpus deseja aplicar o contador. O resultado do contador de freqüência padrão traz diversas informações a respeito das palavras ou expressões do córpus: i) a quantidade de textos pertencentes ao córpus; ii) a quantidade de ocorrências simples (tokens) do córpus; iii) a quantidade de ocorrências simples que aparecem apenas uma vez, bem como, as que aparecem mais de uma vez; iv) a quantidade de "palavras" (lexias complexas) que aparecem no córpus; v) a quantidade de "palavras" que aparecem apenas uma vez, bem como as que aparecem mais de uma vez, e finalmente; e vi) o índice vocabular, que indica a variedade do vocabulário utilizado no córpus.

b) Contador de Freqüência por Palavra ou Expressão: possui a funcionalidade de contar a freqüência de uma palavra ou expressão previamente fornecida pelo usuário. Esse contador é semelhante ao descrito anteriormente, mas, nesse caso, uma palavra ou expressão é requerida como entrada. A palavra ou expressão fornecida pelo usuário pode ser também

---

[15] *Lexia complexa* pode ser entendido como a unidade de significação composta de mais de um token não unidas por meio de hífen. Ex.: ticket refeição, vale transporte, virgem Maria, etc..

[16] Aqui, considera-se palavra complexa as unidades de significação unidas por meio de hífen ou que se constituem pela união de uma seqüência alfabética e outra numérica. Ex.: sem-terra, pára-choque, Largo 13, Pio XI, etc.

uma palavra composta ou lexia complexa. São também dados de entrada o córpus de origem e a "janela" (quantidade de palavras no contexto superior e inferior) onde a mesma aparece.

c) Concordaceador: essa ferramenta tem o objetivo de destacar uma determinada palavra ou expressão no texto onde ela ocorre. O concordanceador implementado no Portal LW possui várias opções que podem ser definidas pelos usuários. Por exemplo, a definição do tamanho do contexto (reduzido e expandido) que dizem respeito, respectivamente, ao tamanho (em caracteres) do segmento e do parágrafo onde a palavra ou expressão aparece, bem como o "nível de sensibilidade", que pode ser: *Igual a, Começando com, Terminando com* e *Contendo* dos mesmos a serem pesquisados no córpus. O resultado da aplicação dessa ferramenta traz todos os trechos (contexto reduzido) nos quais a palavra aparece, sendo que um link sobre a palavra "alvo" leva o usuário aos contextos expandidos. Um concordanceador semelhante é aplicado no córpus Mac-Morpho, com a diferença de que o usuário pode também definir qual etiqueta da palavra ou expressão ele deseja considerar.

Uma importante vantagem das interfaces de ferramentas descritas nesta seção é que todos os resultados podem ser salvos pelo usuário através do link "download do resultado", disponibilizado pela interface. Esta característica oferece maior flexibilidade de navegação e uso de córpus aos usuários, visto que, uma vez aplicadas as ferramentas, os usuários podem, além de se envolver criteriosamente na escolha de seu subcórpus, salvar os seus resultados localmente, o que permite analisá-los posteriormente, i.e., fora do ambiente do portal.

5. Conclusões e Trabalhos Futuros

Quase ao final de 30 meses de pesquisa e desenvolvimento, o LW disponibiliza, de forma gratuita: a) 4 tipos distintos de córpus (Lácio-Ref, Mac-Morpho, Par-C e Comp-C); b) ferramentas de processamento lingüístico-computacional (contador de freqüência, concordanceador e etiquetadores morfossintáticos); e c) Portal com 3 tipos de interface de pesquisa, com ferramentas de base associadas. É, também, um ambiente de navegação dinâmica, didática e, sobretudo, de incentivo ao uso de córpus para os mais diversos tipos de investigação lingüística, uma vez que permite o download completo das amostras dos córpus. É um primeiro passo para um trabalho conjunto de construção de um CNPB.

Embora várias decisões tomadas no projeto do LW ainda estão um pouco distantes dos padrões internacionais (como o XCES) tanto com relação à anotação como à codificação, demos um grande passo em direção à padronização com: a proposta de um rico cabeçalho em XML que traz informações bibliográficas e da tipologia quadripartida; e a anotação explícita da existência de elementos gráficos retirados dos textos. Num possível retorno ao Projeto, espera-se que as limitações na construção e disponibilização de córpus sejam eliminadas: preenchimento com amostras textuais das categorias de gênero e tipo textual, domínio e meio de distribuição não contempladas; estudo e aplicação do balanceamento de córpus; refinamento de ferramentas; associação de novas ferramentas aos córpus e/ou ferramentas já existentes a outros córpus.

Referências

Aluísio, S. M., Pinheiro, G. M., Finger, M., Nunes, M.G.V. and Tagnin, S. E. O. (2003a) "The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation", Corpus Linguistics 2003, Lancaster, UK, Proceedings of Corpus Linguistics 2003. Lancaster: 2003. v. 16, 14-21.

Aluísio, S. M., Pelizzoni, J. M., Marchi, A. R., Oliveira, L. H., Manenti, R. and Maquiafável, V. (2003b) “An account of the challenge of tagging a reference corpus of Brazilian Portuguese”, Lecture Notes on Artificial Intelligence 2721, 110-117.

Aluísio, S. M., Pinheiro, G. M., Manfrim, A. M. P., Oliveira, L. H. M. de, Genovês Jr. L. C. e Tagnin, S. E. O. (2004) “The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools”, LREC 2004. Proceedings of LREC, 2004, Lisboa, Portugal.

Andersen, M. S., Asmussen, H. e Asmussen, J. (2002) “The project of Korpus 2000 going public”, Proceedings of Euralex 2002, 291-299.

Ide, N. e Romary, L. (2003). “Outline of the International Standard Linguistic Annotation Framework.”, Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right, Sapporo, 1-5.

Ide, N., Romary, L., de la Clergerie, E. (2003). “International Standard for a Linguistic Annotation Framework”, Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology, Edmunton.

Ide, N. e Macleod, C. (2001). “The American National Corpus: A Standardized Resource of American English”, Proceedings of Corpus Linguistics 2001, Lancaster UK.

Ide, N. (2000). “The XML Framework and Its Implications for the Development of Natural Language Processing Tools”, Proceedings of the COLING Workshop on Using Toolsets and Architectures to Build NLP Systems, Luxembourg, 5 August 2000.

Ide, N. e Brew, C. (2000). “Requirements, Tools, and Architectures for Annotated Corpora”, Proceedings of Data Architectures and Software Support for Large Corpora. Paris: European Language Resources Association, 1-5.

Ide, N., Bonhomme, P. e Romary, L. (2000). “XCES: An XML-based Standard for Linguistic Corpora”, Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, 825-830.

Ide, N. (1998). “Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora”, Proceedings of the First International Language Resources and Evaluation Conference, Granada, Spain, 463-470.

Santos D. e Sarmento, L. (2003). "O projecto AC/DC: acesso a corpora / disponibilização de corpora", Amália Mendes & Tiago Freitas (orgs.), Anais do XVIII Encontro da Associação Portuguesa de Linguística (Porto, 2-4 de Outubro de 2002), APL, 2003, 705-717.

Santos, D. e Bick, E. (2000) “Providing Internet Acces to Portuguese Corpora: the AC/DC Project”, Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), 205-210. Atenas, 31 May-2 June 2000.

Um Modelo de Identificação e Desambigüização de Palavras e Contextos

Christian Nunes Aranha[1], Maria Cláudia de Freitas[2], Maria Carmelita Pádua Dias[2], Emmanuel Lopes Passos[1]

[1]Departamento de Engenharia – Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) Rio de Janeiro – RJ – Brasil;

[2]Departamento de Letras – Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) Rio de Janeiro – RJ – Brasil;

{chris_ia, emmanuel}@ele.puc-rio.br, {claudiaf, mcdias}@let.puc-rio.br

Abstract. *This paper focus on the interpretation of polisemic words and contexts. It suggests that this interpretation is context-sensitive, regardless of the inner representation of the words themselves. Context is viewed as a cluster formed by a target-word and the words co-ocurring with it. This cluster is achieved by means of statistical treatment of texts as well as graphs representing data obtained in the processing. Preliminary results show that this kind of approach is promising in terms of polisemic words disambiguation.*

Resumo*. Este trabalho aborda a questão da interpretação de palavras e contextos polissêmicos, e sugere que apenas o contexto é necessário para tal interpretação, independente de uma representação interna e autônoma das palavras. O contexto é visto como um aglomerado de palavras coocorrentes com uma palavra-alvo. Esse aglomerado é decorrente de um tratamento estatístico de textos, bem como de grafos que representam os dados obtidos. Os resultados preliminares mostram que este tipo de abordagem é promissor em termos de desambigüização de palavras polissêmicas.*

1. Introdução

Atribuir um significado a uma palavra ou distinguir os diferentes significados de uma palavra polissêmica em um dado contexto são tarefas corriqueiras para qualquer falante de uma língua. Porém, do ponto de vista do Processamento em Linguagem Natural (PLN), a situação não é tão simples. Paradoxalmente, é cada vez mais evidente a necessidade de programas capazes de lidar com a desambigüização de palavras – na recuperação de informações, por exemplo, uma busca por palavra-chave que eliminasse os documentos que trazem esta palavra com um significado não apropriado seria altamente desejável.

O presente trabalho apresenta resultados preliminares provenientes de um projeto que consiste em pesquisar e elaborar modelos relacionados ao Processamento de Linguagem Natural em um único módulo, chamado Cortex. Apresentamos aqui alguns resultados referentes ao processamento de itens lexicais, especificamente à representação automática do significado de palavras inseridas em um contexto. O

princípio subjacente é o de que aspectos das propriedades de uma palavra podem ser capturados por meio de dados estatísticos relativamente às outras palavras que tendem a ocorrer próximas à palavra alvo. Em termos teóricos, buscamos respaldo em teorias vinculadas às correntes pragmáticas do significado – como as de Firth (1957), de inspiração wittgensteiniana -, segundo as quais o significado de uma palavra só pode ser determinado quando inserido em um contexto de uso. O Cortex utiliza uma abordagem puramente estatística da informação lexical. Com um processamento estatístico das palavras de um corpus, o programa é capaz de distinguir os diferentes contextos em que uma determinada palavra pode acontecer, explicitando assim seus possíveis significados e desfazendo ambigüidades.

2. Delimitação do problema: a representação de palavras polissêmicas

O léxico, cada vez mais, vem sendo reconhecido como um dos pontos-chave de programas que visam lidar com PLN. Nesse âmbito, assumem importância fundamental as questões relativas à representação e à aquisição lexicais – de um lado, como representar uma palavra e seu(s) significado(s); e, de outro, como construir um léxico capaz de adquirir novas palavras.

Com relação à representação do significado, a polissemia aparece como um problema teórico, mas não prático. Ou seja, embora de difícil definição em termos teóricos, na prática a polissemia não apresenta qualquer dificuldade – o que Taylor (2003) chama de "o paradoxo da polissemia".

A própria definição tradicional de polissemia – a existência de significados distintos, porém relacionados, em uma mesma palavra – traz consigo questões nada triviais, como qual é a natureza do significado de uma forma lingüística, e o que se entende por relações entre os significados.

A polissemia é um fenômeno inerente a todas as línguas naturais e ela raramente se constitui em um problema de comunicação entre as pessoas. Nenhum falante experimenta qualquer dificuldade para interpretar palavras polissêmicas em seu cotidiano. No entanto, quando se trata relacionar os diversos significados de uma palavra, os falantes apresentam uma grande variação, em relação ao número de acepções bem como a como representá-las. Especialmente no caso de palavras polissêmicas, falantes tendem a discordar quanto às distinções entre significados, e às vezes o mesmo falante pode divergir em suas opiniões sobre o(s) significado(s) de uma palavra.

Um outro exemplo da dificuldade de se lidar com a polissemia é a variação existente, entre dicionários de uma mesma língua, para enumerar e definir significados de uma mesma palavra. Conseqüentemente, propostas de tratamento do significado baseadas em *machine readable dictionaries* (MRD) também costumam apresentar problemas, como demonstram Fillmore & Atkins (2000).

Do ponto de vista computacional, como representar o(s) significado(s) igualmente se mostra um desafio ainda maior, tendo em vista as necessidades de formalização e delimitação necessárias ao meio eletrônico. Taylor resume esse problema, afirmando que

> "a sentence containing *n* words each of which is *m*-times polysemous will in principle have *n* x *m* potential readings. (…)It is not surprising, therefore, that

disambiguation is a major issue en natural language processing." (Taylor 1995:647-648)[1]

Com relação à aquisição lexical, o problema principal consiste em como representar o léxico - um conjunto com um número potencialmente infinito de elementos – de forma a permitir o acréscimo de itens sem comprometer ou modificar o sistema. Outro problema é o fato de novos significados poderem ser incorporados a palavras já existentes (fato comum no caso de terminologias técnicas), o que, de certa forma, nos faz retornar à questão da polissemia.

O interesse na desambigüização de palavras não é recente, e remonta a 1950. (cf *Computational Linguistics* 1998 volume especial sobre desambigüização). Os primeiros trabalhos consistiam em elaborar classificadores "especialistas" que fossem capazes de enumerar os diferentes contextos em que uma palavra pudesse aparecer. Tais classificadores, porém, eram construídos manualmente, o que apresenta um problema para o processamento automático. Como já foi dito, o léxico é um "sistema" aberto: palavras novas são criadas, bem como novos significados para palavras já existentes são cunhados a todo momento. Alimentar manualmente uma base lexical seria um trabalho infindável que, além de tempo, também dependeria de uma vasta mão de obra, o que parece pouco vantajoso. Posteriormente, à medida que *machine readable dictionaries* (MRD) e bases lexicais do tipo WordNet (Fellbaum 1998) se popularizaram, passaram a ser utilizados no fornecimento de informações para a desambigüização automática. Do mesmo modo, porém, a utilização de bases lexicais "prontas" também não parece uma boa solução, pois só há um deslocamento do problema, uma vez, na maioria das vezes, estas são alimentadas manualmente.

Um tratamento realmente automático de dados lexicais, com vistas à interpretação semântica de palavras polissêmicas, pode ser vislumbrado com abordagens estatísticas, como veremos a seguir.

3. Abordagens estatísticas no tratamento lexical

Tentando eliminar o trabalho humano das tarefas de aquisição/ representação lexical, tem-se investido em abordagens estatísticas do léxico, que vêm trazendo resultados promissores e a possibilidade de tratamento de fenômenos como a polissemia (Schütze 1998, Widdows 2002, 2003, Farkas & Li 2002).

As abordagens estatísticas podem ser baseadas em aprendizagem supervisionada e aprendizagem não-supervisionada. No primeiro caso, o processo de desambigüização faz uso de um corpus de treinamento já rotulado. Cada ocorrência de uma palavra ambígua é anotada com um rótulo semântico. Na aprendizagem não-supervisionada, o que está disponível para o treinamento é um corpus não rotulado.

Em termos gerais, modelos estatísticos baseados em coocorrência funcionam da seguinte maneira: a partir de um vasto corpus textual, conta-se, para uma dada palavra-alvo, o número de palavras que aparecem ao seu lado em uma janela de tamanho pré-determinado – por exemplo, 15 palavras. Na etapa seguinte, cada palavra é representada por meio das freqüências cumulativas das ocorrências no escopo da janela. Palavras

[1] "Uma sentença que contenha *n* palavras, cada uma delas *m* vezes polissêmica terá em princípio n x m leituras potenciais. (...) Não é de surpreender, então, que a desambigüização seja uma importante questão no processamento de linguagem natural". (Tradução dos autores)

com significados similares tenderão a ocorrer em contextos similares e palavras polissêmicas tenderão a ocorrer em contextos diferentes.

Subjacente a esses modelos, está a idéia de que o significado de uma palavra corresponde ao seu padrão de uso, e não ao significado considerado autonomamente. Porém, muitas vezes a decisão de não considerar o significado propriamente dito – ou intrínseco – das palavras é tomada por praticidade, pois, como diz Schütze, "(...) *providing information for sense definitions can be a considerable burden*."(1998: 97).[2] Segundo o autor (Schütze 1998), para se definir o significado "verdadeiro" das palavras – o que ele chama de etiquetagem de significados (*sense labeling*) –, é necessário se levar em conta uma fonte externa de conhecimento, que pode ser tanto um dicionário, um corpus bilíngüe, *thesauri* ou conjuntos de treinamento de etiquetados manualmente.

Schütze, assim, aponta para a dificuldade de se chegar ao significado "verdadeiro" de uma palavra; entretanto, ele não nega a sua existência. Do mesmo modo, Widdows (2003), que também apresenta um modelo de aquisição e desambigüização lexical baseado em informação contextual, afirma que o significado pode ser descrito de forma "clara, flexível e acurada", através de um pensamento científico cuidadoso e de investigação empírica. Ainda segundo Widdows (2003), métodos estatísticos, embora tenham trazido enormes contribuições, apenas adivinham o significado das palavras.

Embora o Cortex também desconsidere o significado propriamente dito, ou intrínseco, das palavras, o faz motivado teoricamente. No âmbito de uma teoria de inclinação pragmática como a de Firth, o significado de uma palavra é compreendido justamente como decorrência das suas relações com o contexto. Seguindo a linha wittgensteiniana, Firth afirma que "you shall know the meaning of a word by the company it keeps" (1957: 194-6); e, de acordo com Cruse (1986), "o significado de uma palavra é constituído por suas relações contextuais". Ou seja, parte-se do pressuposto de que não há significado fora de um contexto. No caso específico do processamento realizado pelo Cortex, "contexto" corresponde estritamente ao ambiente lingüístico em que uma palavra pode ocorrer, e nada mais além disso. De forma mais específica, contexto corresponde a uma janela cujo limite é o ponto final.

No Cortex, o significado é compreendido como uma rede de relações entre as palavras; o significado de uma palavra *p* é determinado pelas relações entre *p* e as outras palavras que coocorrem com *p*. Especificamente, a cada significado de *p* corresponde uma rede de relações diferente. Assim, por exemplo, a palavra *ataque* pode vir numa relação com *jogador* e *futebol*, em que é possível depreender o seu significado inserido em uma situação de *esportes*. Em outro contexto, pode vir acompanhada de *bombas* e *terrorismo*, o que reflete o significado de *agressão*, e ainda pode aparecer coocorrendo com *sintoma* e *medicamento*, incorporando o significado de *acesso de doença*.

Tomamos também como ponto de partida que mesmo as palavras que não são tradicionalmente consideradas polissêmicas precisam ser "desambiguizadas", pois apenas o contexto de uso faz refletir a interpretação a ser tomada pela palavra em questão. Uma palavra como *jogador*, por exemplo, pode parecer tanto em um contexto que dirija a significação para *jogador de vôlei* quanto em um contexto que dirija a

[2] (...) fornecer informações para definições de sentido pode ser uma tarefa considerável". (Tradução dos autores)

significação para *jogador de futebol*. Essa característica fica especialmente evidente em PLN, uma vez que todas as palavras são potencialmente ambíguas (polissêmicas ou homônimas) e só o contexto desfaz a ambigüidade.

4. O Cortex

A abordagem do Cortex toma como inspiração o modelo de Schütze (1998), segundo o qual o significado de uma palavra ambígua pode ser distinguido a partir da análise dos seus padrões de contextualização. Neste modelo, tanto os significados quanto os contextos de uso de uma palavra ambígua são representados como direções em um espaço vetorial, e um contexto é atribuído a um significado quando ambos possuem a mesma direção. A idéia básica é que quanto mais vizinhos em comum duas palavras tiverem, mais similares elas serão; e quanto mais palavras similares aparecerem em dois contextos, mais similares os dois contextos serão. O modelo compreende duas etapas: treinamento e desambigüização. Na primeira etapa, em um corpus de treinamento não-rotulado, acontece a contagem da freqüência de coocorrência entre as palavras. Nesse momento são calculados os vetores das palavras, os vetores de contexto e os vetores de significado. Todos os contextos de uma palavra ambígua são coletados no corpus de treinamento. Numa etapa posterior, já no corpus de testagem, a partir das informações coletadas na etapa de treinamento, é possível desambiguizar uma determinada palavra-alvo.

No Cortex, assim como verificado em Schütze (1998), a hipótese subjacente à desambigüização é a de que o significado pode ser caracterizado em termos de padrões de contextualização. Por isso, no processamento realizado no Cortex também não há rótulo ou etiquetagem das palavras, ou seja, não há um valor intrínseco para cada palavra. A forma de se chegar ao "significado" é através das relações de coocorrência entre as palavras.

O Cortex contém um algoritmo estatístico que extrai conhecimento sobre o contexto das palavras em um corpus não rotulado. O algoritmo é aplicado diretamente ao corpus de testagem (no caso aqui apresentado, composto por dois meses de notícias de jornal).

A discriminação do contexto ocorre da seguinte maneira: no escopo de uma janela cujo limite é o ponto final[3], conta-se, ao longo do corpus, a quantidade de vezes que uma palavra coocorreu com todas as outras palavras. É realizado, então, um teste de hipótese para determinar a significância da relação entre duas palavras, isto é, se elas ocorreram ao acaso ou não. Um grafo é formado tomando-se como nós as palavras e arestas de todas as relações significativas entre os nós. Palavras funcionais não são computadas neste processo: um banco lexical com uma lista destes itens trata de eliminá-los a fim de reduzir o esforço computacional. Esta escolha deve-se ao fato de que palavras funcionais apresentam alta freqüência de ocorrência e também de coocorrência, o que as torna candidatas a estarem presentes no contexto por mero acaso.

A partir de uma palavra-alvo *p*, é utilizado um algoritmo que busca as palavras mais relacionadas a p e que têm, simultaneamente, uma grande quantidade de ligações

[3] Segundo Manning & Schütze (1999), cerca de 90% dos sinais de ponto de um texto correspondem realmente a pontos finais; logo, é razoável adotar o ponto como limite para tratamento estatístico de textos.

entre si. Essas palavras irão constituir um aglomerado, que, de certa forma, pode ser considerado um campo semântico. Palavras que contêm muitas ligações são consideradas fracas, e são dispensadas logo no início, já que suas ligações acabam sendo pouco representativas. Uma mesma palavra pode pertencer a diferentes aglomerados, o que seria indicativo de sua polissemia. Como já mencionado, não apenas palavras tradicionalmente consideradas polissêmicas apresentam diferentes aglomerados (isto é, diferentes contextos). Uma palavra como *jogador*, por exemplo, pode aparecer tanto no contexto vôlei como no contexto futebol. Quanto mais aglomerados forem detectados, mais refinada será a distinção entre as palavras.

Além da possibilidade de desambigüização de qualquer palavra, e não apenas das classificadas como ambíguas, isto é, aquelas para as quais foram encontrados dois contextos distintos no corpus de treinamento, e da ausência de um número pré-determinado de contextos, o Cortex difere do modelo de Schütze por ser uma abordagem híbrida que utiliza, além de estatística, otimização por grafos. Nesse aspecto, o modelo se aproxima da abordagem de Widdows (2003) e Widdows & Dorow (2002), a qual busca, através de grafos, demonstrar relações semânticas entre as palavras. Após a montagem do grafo, basta uma condição inicial, ou uma palavra *p*, para que o sistema encontre automaticamente todos os diferentes contextos em que *p* aparece – ou seja, todos os seus "significados". Esses resultados podem variar em função de alguns parâmetros que devem ser configurados antes da busca. São eles:

N: quantidade máxima de sentidos a serem procurados = 100
A: quantidade de palavras armazenadas por contexto inicial = 10
L: limite permitido de ligações que uma palavra pode ter para participar de um contexto = 100%
S: fator de similaridade para unir dois contextos = começa com 90% e diminui até 70%

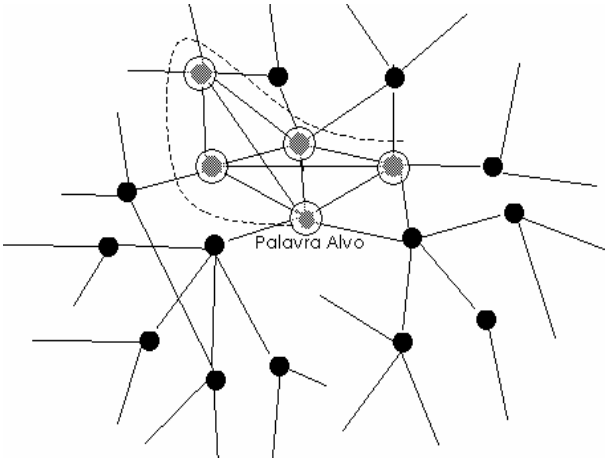


Figura 1. Criação dos contextos

O algoritmo de *clustering* é iterativo e funciona como uma aranha movimentando-se em uma teia. Inicialmente posicionada sobre a palavra-alvo *p*, a "aranha" passeia pelo grafo de acordo com as ligações mais fortes. É criado o "contexto 1", partindo da palavra mais forte, e, a cada passo, testamos se o novo nó (palavra) tem uma quantidade de ligações suficiente (L) para entrar no contexto. Em caso afirmativo,

a aranha adiciona a palavra ao contexto 1 e continua a passear na teia para a próxima palavra, e assim sucessivamente. Em caso negativo, o contexto 1 é fechado e é criado um novo contexto ("contexto 2"). O mesmo procedimento é realizado partindo da segunda palavra de relação mais forte. Os contextos vão sendo criados até acabarem as palavras.

Como o número de contextos encontrados inicialmente foi bastante grande, foi necessário limitar a procura em N contextos, e criou-se uma 2ª etapa para realizar um enxugamento dos contextos encontrados. Assim, uma vez detectados todos os contextos possíveis, o programa passa a re-agrupar da seguinte forma: primeiramente aglutina os conjuntos com similaridade de 90%, e, com esse resultado, aglutina novamente até não ser mais possível formar nenhum conjunto. Em seguida, o valor de S é diminuído para 80% e todo o procedimento se repete. Em uma última etapa, o valor de S cai para 70%.

O resultado da aplicação dos algoritmos é uma lista de palavras representantes de um contexto. Nos quadros 1 e 2, abaixo, estão os resultados de busca para as palavras *título* e *prova*.

Quadro 1: resultado de processamento do Cortex para a palavra *título*

1.		**2.**		**3**.
dívida	histórico	brasileira	Brasil	Milan
brasileira	novo	principal	vitória	sexto
Valor	US	brasileiro	time	europeu
principal	Brasil	final	estréia	
externa	subia	Milan	Roma	
face	consecutivo	Manchester	campeão	
Bond	histórica	domingo	espanhol	
alta	tarde	Lancepress	vaga	
cotado	lucros	competição	Federer	
negociado	imposto	conquistar	Santos	
final	inédito	Campeonato	inglês	
opera	internacionais	turno	gols	
valorização	positiva	conquistou	Fábio	
recorde	registrada	conquista	mantém	
cotação	segue	disputa	Emanuel	
dia	cai	seleção	conquistado	
risco	mantém	feira	Mantilla	
Real	disco	Liga	derrotar	
exterior	queda	jogos	casa	
tendência	investidores	mundial		
feira				

Quadro 2: resultado de processamento do Cortex para a palavra *prova*

1.	**2.**	**3**.	**4**.	**5**.	**6**.
Venceu domingo lugar brasileiro GP vencedor	piloto corrida brasileiro Indianápolis dia	balas coletes bala colete	especial domingo americano brasileiro Indianápolis Pan Americano agosto	pegada importando Americaninha pedofilia Memorando gente	exame legítima Jamilly

No Quadro 1, estão os resultados do processamento realizado tendo a palavra *título* como alvo. O sistema encontrou três contextos para ela. A partir das palavras presentes no contexto 1, é possível perceber que título está sendo utilizado em um contexto de economia. As palavras coocorrentes – *cotação*, *dívida, externa*, *investidores* e *lucros*, entre outras –sugerem tal interpretação. Já o contexto 2 atribui a *título* o significado de *prêmio em uma competição esportiva*, como sugerem as palavras *campeão*, *campeonato*, *competição*, *conquista*, *jogos*, *time*, entre outras. Por fim, o contexto 3 parece ser um refinamento do contexto anterior: trata-se de uma competição européia, como sugerem as palavras *europeu* e *Milan.*

Pelo Quadro 2, é possível observar que o sistema foi capaz de distinguir 6 contextos de uso de *prova*. A partir das palavras presentes no contexto 1, é possível atribuir a *prova* o significado de *competição*. Já o contexto 2 pode ser entendido como uma subdivisão (refinamento) do contexto 1: não se trata de uma competição qualquer, mas de uma *competição de corrida*, possivelmente de carros, como indicam as palavras coocorrentes *corrida* e *piloto*. No contexto 3*, prova* adquire o valor de *resistente a – à prova de balas*, no caso. A identificação do contexto 4 não é tão evidente, mas a presença da palavra *Pan* sugere que também se trate do contexto *competição*. Por fim, os contextos 5 e 6 parecem atribuir à prova o significado de *indício*: no contexto 5, as palavras coocorrentes *pedofilia* e *pegada* levam a esta interpretação; no contexto 6, *exame* e *legítima* também sugerem que *prova* está sendo interpretada como *indício*.

Um detalhe que chama a atenção em ambos os quadros é o grande número de nomes (substantivos e adjetivos), em oposição a verbos, presentes nos aglomerados. No grupo *prova*, por exemplo, das 32 palavras utilizadas na caracterização, apenas duas são verbo (e, assim mesmo, uma delas é uma forma nominal de verbo). No grupo *título*, das 83 palavras, apenas 8 são indubitavelmente verbos, o que pode ser indicativo da força da classe nominal na caracterização de contextos.

5. Considerações Finais e Direcionamentos Futuros

Apresentamos aqui os resultados preliminares do Cortex, um processador de linguagem natural. Tais resultados são relativos à atribuição de significado às palavras, e sua conseqüente desambigüização. O sistema é inspirado no modelo de Schütze (1998) e é capaz de identificar todos os contextos relativos à palavra-alvo que aparecerem no corpus.

Algumas questões permanecem em aberto. Uma delas é em que medida é vantajosa uma definição altamente refinada de uma série de contextos, como é possível observar na desambigüização de *prova*. Talvez não haja uma única resposta, e o grau de vantagem dependa diretamente dos objetivos do usuário.

De forma geral, os resultados, embora significativos, ainda precisam de ajustes.ë o caso, por exemplo, de contextos que parecem incluídos em outros, como aconteceu com a palavra *título*, em que o contexto 3 é um subconjunto do contexto 2. Algumas melhorias já estão sendo incorporadas para possibilitar um "enxugamento" nos resultados. Por exemplo, a duplicação de palavras decorrentes da flexão de plural nos nomes será eliminada. Assim, no contexto 3 de *prova*, por exemplo, *bala/balas* e *colete/coletes* darão lugar a somente uma instância de cada palavra: no resultado final aparecerá apenas *bala* e *colete*. Do mesmo modo a presença de formas flexionadas de um verbo será eliminada no resultado final. No contexto 2 de *titulo*, *conquistar* e *conquistou* darão lugar a uma instância apenas, possivelmente *conquistar*.

Outro ajuste diz respeito aos nomes próprios e compostos, que atualmente são considerados duas palavras distintas, como é o caso de *Estácio de Sá*, por exemplo, que apareceu em outro experimento como duas palavras (*Estácio* e *Sá*), quando na verdade é um nome próprio composto. Pretende-se refinar o processamento para incluir expressões compostas. Além disso, combinações de duas ou três palavras com um padrão muito alto de coocorrência passarão a ser consideradas apenas um item lexical. No contexto 4 do grupo prova, muito provavelmente as palavras *Pan* e *americano* se juntariam para formar o item *Pan americano*, e no grupo 1 de *título*, o mesmo aconteceria com *dívida* e *externa*, que passaria a ser considerada *dívida externa*.

Ajustes servirão também para definir melhor, ou eliminar, contextos pouco claros, como é o caso dos contextos 5 e 6 de *prova*.

Ainda assim, acreditamos que os primeiros resultados conseguidos pelo Cortex apontam para um possível caminho para a utilização de dados estatísticos para desfazer automaticamente a ambigüidade de palavras em contexto.

Referências Bibliográficas

Cruse, D. (1986) Lexical Semantics. Cambridge: CUP.

Farkas, I.& Li, P. (2002) "Modeling the development of lexicon with a growing self-organizing map", Proceedings of the 6th Joint Conference on Information Sciences, Research Triangle Park, NC, p. 553-556.

Fillmore, C. & Atkins, B. (2000) "Describing polisemy: the case of 'Crawl' ", In: Ravin & Leacock (eds.). Polysemy. Theoretical and computational approaches. Oxford: Oxford University Press.

Firth, J. R. (1957). Papers in Linguistics – 1934-1951. Oxford: Oxford University Press.

Manning, C & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press.

Schütze, H. (1998). "Automatic Word Sense Discrimination", Computational Linguistics, 24(1), 97-123.

Schütze, H & J. Pedersen.(1995). “Information retrieval based on word senses”, Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, EUA, p. 161-175.

Taylor, J. (2003). “Polisemy’s paradoxes”, Language Sciences (25) 637-655.

Widdows, D. (2003). “A Mathematical Model for Context and Word-Meaning”, Fourth International and Interdisciplinary Conference on Modeling and Using Context, Stanford, California, June 2003, pp. 369-382.

Widdows, D & Dorow, B. (2002). “A Graph Model for Unsupervised Lexical Acquisition”, Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, p.1093-1099.

Identificação de Expressões Anafóricas e Não Anafóricas com Base na Estrutura do Sintagma

Sandra Collovini, Rodrigo Goulart, Renata Vieira

Programa Interdisciplinar de Pós-Graduação em Computação Aplicada – PIPCA

Universidade do Vale do Rio dos Sinos (UNISINOS)

Av. Unisinos, 950 – 93.022-000 – São Leopoldo, RS – Brasil

{sandrac, rodrigo, renata}@exatas.unisinos

Abstract. *One of the problems in anaphora resolution is to identify which expressions are anaphoric and which are non anaphoric. In this work a group of heuristic to identify the expressions as non anaphoric, the implementation of these heuristic in an environment for anaphora resolution (ART - Anaphor Resolution Tool) and an evaluation of the obtained results is presented.*

Resumo. *Um dos problemas da resolução de anáforas é identificar quais expressões são anafóricas e quais são não anafóricas. Neste trabalho um conjunto de heurísticas para identificar as expressões não anafóricas, a implementação destas heurísticas em um ambiente para a resolução de anáforas (ART - Anaphor Resolution Tool) e uma avaliação dos resultados obtidos são apresentados.*

1. Introdução

Este trabalho trata da resolução de expressões anafóricas em textos da Língua Portuguesa, mais especificamente descrições definidas. Chamamos descrições definidas os sintagmas nominais iniciados por artigo definido (o, a, os, as). Uma expressão é considerada anafórica quando se refere a uma entidade previamente referenciada no texto por meio de outra expressão. A resolução de anáforas consiste em encontrar um antecedente para os sintagmas nominais.

A identificação de expressões anafóricas, que se referem à mesma entidade, é importante em diversas aplicações de Processamento de Linguagem Natural, por exemplo, em sumarização automática, extração de informação, recuperação de informação, tradução automática, classificação de textos, entre outros.

Trabalha-se com descrições definidas, pelo fato de ocorrerem em grande quantidade nos textos do tipo de corpus estudado [Vieira et al., 2002]. Além disso, existem vários trabalhos sobre resolução de anáforas pronominais, mas não existem muitos trabalhos que tratem especialmente as relações de co-referência entre as descrições definidas.

Estudos recentes mostram que as descrições definidas além de ocorrerem em grande número, somente em 50% dos casos são consideradas expressões anafóricas. Por isso, consideramos importante o desenvolvimento de heurísticas para identificação de descrições definidas não anafóricas no processo de resolução dessas expressões.

Neste trabalho apresenta-se a implementação e a avaliação de heurísticas no ambiente ART (*Anaphor Resolution Tool*) [Gasperin et al., 2003] para a tarefa de identificar descrições definidas não correferentes[1], ou seja, expressões não anafóricas, com base na estrutura sintática do sintagma nominal, de acordo com estudos prévios feitos para a Língua Inglesa [Vieira, 1998].

O trabalho encontra-se assim organizado: na seção 2, são apresentadas algumas considerações sobre resolução de anáforas e um estudo do sintagma nominal. Na seção 3, é mostrada a análise de corpus. Na seção 4, uma visão geral do Ambiente de Desenvolvimento ART é dada e as heurísticas para a classificação automática são apresentadas em detalhadamente. Por fim, na seção 5 são avaliados os resultados juntamente com as considerações finais.

2. Resolução de Anáforas

A tarefa de resolução de anáforas consiste na identificação de um antecedente textual importante na interpretação de uma expressão, como é ilustrado no exemplo a seguir: *"O Eurocenter oferece cursos de Japonês na bela cidade de Kanazawa, tanto para iniciantes quanto para aqueles com conhecimento avançado da língua. Os cursos têm quatro semanas de duração".*

Devido à complexidade da tarefa de encontrar um antecedente, somado ao fato de nem todas as expressões serem anafóricas (principalmente as descrições definidas), a comunidade vem propondo que parte do processo de resolução consiste em diferenciar sintagmas nominais entre anafóricos e não anafóricos [McCarthy and Lehnert, 1995; Bean and Riloff, 1999; Cardie and Wagstaff,1999; Vieira and Poesio, 2000; Soon et al., 2001; Muller et al., 2002; Ng and Cardie, 2002a; Ng and Cardie, 2002b; Uryupina, 2003]. Para isso, uma análise do sintagma nominal do português foi realizada para adaptar heurísticas do inglês na identificação de sintagmas nominais não anafóricos.

2.1. Estudo do Sintagma Nominal

Os sintagmas são formados por vários elementos que constituem uma unidade significativa dentro da sentença, além de manterem entre si relações de dependência e de ordem [Silva and Koch, 1989]. Estes elementos podem ser uma única palavra ou um conjunto de palavras. Os sintagmas desempenham uma função na sentença e combinam-se em torno de um núcleo. A classificação do sintagma é dependente do seu núcleo, por exemplo, quando o núcleo for um nome o sintagma é classificado como sintagma nominal. Conforme Perini (2003), o sintagma nominal possui uma estrutura bastante complexa, pois é possível distinguir dentro do sintagma nominal várias funções sintáticas. O núcleo do sintagma nominal pode ser um nome (comum ou próprio) ou um pronome (pessoal, demonstrativo, indefinido, interrogativo ou possessivo). O sintagma nominal pode também ser constituído por determinantes e/ou modificadores, sendo que os modificadores antecedem ou sucedem o núcleo, enquanto os determinantes apenas o antecedem [Miorelli, 2001].

Um sintagma nominal pode ser classificado como uma expressão anafórica ou não anafórica dependendo da sua relação de co-referência no discurso. As expressões

[1] Expressões correferentes são diferentes expressões invocando o mesmo referente.

são ditas anafóricas quando fazem referência a uma entidade introduzida no texto. As anáforas podem ser pronominais, definidas, indefinidas ou demonstrativas.

Um sintagma nominal não anafórico, introduz uma nova entidade no modelo discursivo. Geralmente ocorre no início do texto com descrições indefinidas, por exemplo, *"Uma instituição social"* ou com descrições definidas complexas, por exemplo, *"O quilômetro 430 da rodovia Assis Chateau Briand"*.

Nesse trabalho, o foco dos estudos são as descrições definidas. Estudam-se as descrições definidas segundo a classificação apresentada em Vieira (1998):

1. Anafóricas Diretas: são antecedidas por uma expressão que possui o mesmo nome-núcleo e refere-se à mesma entidade no discurso, por exemplo, "*Comprei um sapato. O sapato é confortável"*.
2. Anafóricas Indiretas: são antecedidas por uma expressão que não têm o mesmo nome-núcleo do seu antecedente. Assim, o núcleo pode ser um sinônimo do antecedente ou mesmo uma elipse, referindo-se à mesma entidade já introduzida no discurso, por exemplo, "*Comprei um apartamento. A moradia fica perto daqui"*.
3. Anafóricas Associativas: introduzem um referente novo no discurso, mas que tem uma relação semântica com algum antecedente já introduzido. Assim, a descrição definida tem seu significado ancorado em um referente, por exemplo, "*Ganhei uma rifa. O número sorteado foi o 100"*.
4. Não Anafóricas: são aquelas que introduzem um novo referente no texto que não se relaciona com nenhum antecedente no discurso. Assim, não possui uma âncora para se apoiar semanticamente, por exemplo, "*O radialista da Rádio Globo Washington Rodrigues"*.

3. Análise de Corpus

O corpus utilizado nesse estudo constitui-se de um extrato do corpus NILC[2], formado por 10 textos jornalísticos retirados da Folha de São Paulo, escritos em português do Brasil. Cada documento é um arquivo texto (formato ASCII) com tamanho entre 1 Kbyte e 6 Kbytes, com um mínimo de 41 termos e um máximo de 895 termos.

O corpus estudado foi anotado sintaticamente. Para obter a análise das sentenças do corpus, utilizou-se o analisador sintático PALAVRAS[3] descrito em Bick (2000), uma ferramenta robusta para a análise sintática do português. A partir da saída do analisador sintático a ferramenta Xtractor[4] descrita em Gasperin et al. (2003) gera três arquivos XML. O primeiro arquivo é o *arquivo de Words,* Figura 1; o segundo é o arquivo com as categorias morfossintáticas (*POS – Part of Speech*), Figura 2; e por fim, o terceiro é o arquivo com as estruturas sintáticas das sentenças representadas por *chunks*. Um *chunk*

[2]Núcleo Interinstitucional de Lingüística Computacional. Disponível em http://www.nilc.icmp.usp.br/nilc

[3] O analisador PALAVRAS faz parte de um grupo de analisadores sintáticos do projeto VISL (Visual Interactive Syntax Learning), do Institute of Language and Communication da University of Southern Denmark Disponível em: http://visl.sdu.dk/visl/pt/parsing/automatic/

[4] A Ferramenta Xtractor engloba a análise do corpus a partir do analisador sintático PALAVRAS, o tratamento da saída desse analisador, com a geração de três arquivos XML.

pode possuir sub-elementos *chunks* com informações das sub-estruturas das sentenças, Figura 3.

```
<words>
.....
<word id="word_69">o</word>
<word id="word_70">radialista</word>
<word id="word_71">de</word>
<word id="word_72">a</word>
<word id="word_73">Rádio_Globo_Washington_Rodrigues</word>
.....
</words>
```

Figura 1. Arquivo de Words

```
<words>
.....
<word id="word_73">
<prop canon=
"RádioRádio_Globo_Washington_RodriguesGlobo_Washington_Rodrigues"
gender="M" number="S"/>
</word>
.....
```

Figura 2. Arquivo das Categorias Morfossintáticas

```
<text>
<paragraph id= "paragraph_1">
.....
<sentence id="sentence_7" span="word_69..word_96">
<chunk id="chunk_95" ext="sta" form="fcl"
       span="word_69..word_95">
  <chunk id="chunk_96" ext="subj" form="np"
         span="word_69..word_70">
  <chunk id="chunk_97" ext="n" form="art" span="word_69">
  </chunk>
.....
```

Figura 3. Arquivo de Chunks

Nesse estudo, os atributos dos *chunks* serão utilizados para a implementação das heurísticas no Ambiente ART (seção 4). As informações de interesse dos *chunks* são:

- Atributo *ext*: representa a função do *chunk,* por exemplo, sentença ou enunciado (*ext=sta*); sujeito (*ext=subj*); núcleo (*ext=h*).
- Atributo *form*: representa a forma do *chunk,* tais como: cláusula finita (*form=fcl)*; sintagma nominal (*form=np)*; substantivo (*form=n*).

Depois da anotação sintática automática, o corpus foi analisado manualmente em relação a co-referência. A anotação manual consiste em duas etapas. Em um primeiro momento, são anotadas as descrições definidas, considerando-se que uma

descrição definida pode conter outras descrições definidas, por exemplo, *"A lista do banqueiro do jogo do bicho", "o banqueiro do jogo do bicho", "o jogo do bicho"*. Em um segundo momento, as descrições definidas são classificadas como anafóricas e não anafóricas.

Para a anotação manual do corpus, utilizou-se a ferramenta MMAX (*Multi-Modal Annotation in XML*) [Müller and Strube, 2000], específica para anotação de corpus. Essa ferramenta utiliza o *arquivo de Words*, gerado pela ferramenta Xtractor que contém todas as palavras do corpus associadas a um identificador (atributos *id* da Figura 1). Ela também utiliza um segundo arquivo que contém a estrutura do corpus (parágrafos, sentenças, cabeçalhos, etc), ilustrado na Figura 4.

```
.....
<paragraph>
      <sentence id="sentence_1" span="word_1..word_8"/>
      <sentence id="sentence_2" span="word_9..word_23"/>
</paragraph>
.....
```

Figura 4 . Arquivo da Estrutura

O resultado do processo de anotação no MMAX é um arquivo que contém a anotação de co-referência. As marcações são codificadas como elementos *markable*, cujo atributo *span* indica as palavras que formam a expressão, o atributo *pointe*r indica o identificador do antecedente. Além destes, outros atributos podem ser especificados pelo pesquisador. Para esse estudo, acrescentou-se o atributo *classification* que corresponde à classificação anafórica da expressão (Figura 5).

```
.....
<markable>
      <markable id="markable_1"
      pointer=" "
      span="word_3..word_4"
      classification="non_anaphoric"/>
</markable>
.....
```

Figura 5. Arquivo de Marcações

4. Heurísticas para identificação de descrições definidas não anafóricas

ART é uma ferramenta para resolução de expressões anafóricas, onde o processo de resolução das anáforas é baseado em heurísticas. A ferramenta é desenvolvida em Java e os dados de entrada e saída utilizam a linguagem de marcação XML.

A arquitetura da ferramenta é baseada em "*pipes & filters*", constituindo-se de um conjunto de três passos (baseados na anotação manual) com uma ou mais tarefas codificadas através de folhas de estilo XSL[5] (*eXtensible Stylesheet Language*). As heurísticas utilizam informações dos textos analisados e são implementadas com folhas de estilos XSL.

[5] Linguagem Desenvolvida pelo W3C (world Wide Web Consortium) disponível em: http://www.w3.org/Style/XSL/

Nesse estudo, testamos algumas heurísticas para identificar as descrições definidas não anafóricas com base na estrutura do sintagma. Entre as heurísticas que serão apresentadas, a heurística 1, 2 e 3 foram elaboradas com base nos estudos da Língua Inglesa detalhado em Vieira (1998) e adaptadas para a Língua Portuguesa. Já as heurísticas 4, 5, e 6 foram construídas a partir da análise das características morfossintáticas das descrições definidas do corpus anotado estudado.

Heurística 1: expressão acompanhada de um sintagma preposicional, pós-modificador (restritivo), por exemplo, *"A tarde de ontem"*. Um pós-modificador restritivo sucede o núcleo restringindo-o. Um modificador restritivo permite que o referente seja identificado através da informação do modificador que especifica a informação do núcleo. Procura-se a existência de um sintagma preposicional no *chunk* da descrição definida, ou seja, um filho desse *chunk* com o atributo *form* igual a "*pp*". A Figura 6 ilustra o *span* "word_200..word_203" que corresponde a *"a tarde de ontem"*.

```
.....
<chunk id="chunk_277" ext="p" form="np"
       span= "word_200..word_203">
.....
  <chunk id="chunk_280" ext="n" form="pp"
         span="word_202..word_203">
.....
</chunk>
.....
```

Figura 6. Trecho do Arquivo de Chunks

Heurística 2: expressão constituída de construções de apostos, por exemplo, *"O prefeito de Gravataí, Daniel Luiz Bordignom"*. O aposto é um sintagma composto, com uma expressão adjacente que o explica ou especifica. O aposto pode vir separado por vírgulas ou depois de dois pontos. No corpus estudado, há construções de apostos como no exemplo acima em que o aposto *"Daniel Luiz Bordignom"* é uma explicação sobre "*o prefeito de Gravataí"*. Nessa heurística analisa-se a estrutura sintática do *chunk,* buscando-se por uma construção de aposto, ou seja, um filho com o atributo *ext* igual a "*app*". A Figura 7 ilustra o *span* "word_49..word_54" que corresponde a *"o prefeito de Gravataí, Daniel_Luiz_Bordignom"*.

```
.....
<chunk id="chunk_71" ext="subj" form="np"
       span=" word_49..word_54">
.....
  <chunk id="chunk_78" ext="app" form="prop" span=" word_54">
.....
</chunk>
.....
```

Figura 7. Trecho do Arquivo de Chunks

Heurística 3: expressão acompanhada de uma cláusula relativa, por exemplo, *"O texto que deve ser assinado pelos jornalistas"*. Nessa heurística, procura-se a existência de uma cláusula relativa, isto é, um irmão desse *chunk* que possua o atributo *form* igual a "*pron_indp*". A Figura 8 ilustra o *span* "word_100..word_108" que corresponde a *"o texto que deve ser assinado por os jornalistas"*.

```
.....
<chunk id="chunk_152" ext="subj" form="np"
       span="word_100..word_108">
.....
  <chunk id="chunk_161" ext="subj" form="pron_indp"
         span="word_105">
.....
</chunk>
```

Figura 8. Trecho do Arquivo de Chunks

Como neste trabalho os antecedentes não estão sendo considerados, apenas a estrutura do sintagma, adicionamos algumas restrições às heurísticas relacionadas a nomes próprios utilizadas anteriormente para o inglês (4 e 5).

Heurística 4: expressão com o núcleo sendo um nome próprio composto, por exemplo, "*A Rádio Globo Washington Rodrigues*". No corpus estudado, por tratar-se de textos jornalísticos, são relatadas informações sobre locais, eventos, pessoas, empresas importantes da atualidade, sendo que uma característica observada nesses textos é a presença de nomes próprios compostos, ou seja, nomes próprios formados por dois ou mais elementos que geralmente introduzem um novo referente no discurso. Para tratar desses casos, busca-se o núcleo dessa estrutura, ou seja, o filho desse *chunk* que possua o atributo *ext* igual a "*h*" e a forma de nome próprio, isto é, o atributo *form* igual a "*prop*". A Figura 9 ilustra o span "word_72..word_73" correspondente a *"a Rádio_Globo_Washington_Rodrigues"*.

```
.....
<chunk id="chunk_100" ext="p" form="np"
       span="word_72..word_73">
.....
  <chunk id="chunk_102" ext="h" form="prop" span="word_73">
</chunk>
.....
```

Figura 9. Trecho do Arquivo de Chunks

Heurística 5: expressão acompanhada de um nome próprio, por exemplo, *"O delegado Elson Campelo"*. No corpus estudado, uma característica observada nos textos é a construção de descrições definidas com núcleo sendo um nome comum (substantivo comum), seguido de um nome próprio especificando esse núcleo e geralmente tratando-se de um novo referente no discurso. Para resolver esses casos, analisa-se a estrutura do *chunk* , localizando o seu núcleo, ou seja, o filho desse nodo que possua o atributo *ext* igual a "*h*" e a forma de nome comum (substantivo comum), isto é, o atributo *form* igual a "*n*". Em seguida, verifica-se a presença de um nome próprio, isto é, um irmão desse *chunk* que possua o atributo *form* igual a "*prop*". A Figura 10 ilustra o *span* word_186..word_188 correspondente a *"o delegado Elson_Campelo"*.

```
.....
<chunk id="chunk_258" ext="acc" form="np"
       span="word_186..word_188">
.....
  <chunk id="chunk_260" ext="h" form="n" span="word_187">
  <chunk id="chunk_261" ext="n" form="prop" span="word_188">
</chunk>
.....
```

Figura 10. Trecho do Arquivo de Chunks

Identificamos na análise do corpus um outro tipo de pós-modificador restritivo freqüente, o sintagma adjetival.

Heurística 6: expressão acompanhada de sintagma adjetival, pós-modificador (restritivo), por exemplo, *"Os momentos mais difíceis de minha carreira"*. Um pós-modificador pode se configurar como um sintagma adjetival, que possui como núcleo um adjetivo. Para essa heurística, verifica-se a presença de um sintagma adjetival nessa estrutura, ou seja, o filho desse *chunk* que possua o atributo *form* igual a "*ap*". A Figura 11 ilustra o *span* "word_22..word_28" que corresponde a *"os momentos mais difíceis de minha carreira"*.

```
.....
<chunk id= "chunk_31" ext="p" form="np"
       span= "word_22..word_28">
.....
  <chunk id="chunk_34" ext="n" form="ap"
         span= "word_24..word_28">
.....
</chunk>
.....
```

Figura 11. Trecho do Arquivo de Chunks

De posse das heurísticas desenvolvidas, é possível automatização do processo de resolução de anáforas.

5. Avaliação

Na seção anterior foi apresentado um conjunto de heurísticas para identificar as descrições definidas não anafóricas e a implementação dessas heurísticas no Ambiente ART. Para analisar os resultados, é necessário comparar os resultados da aplicação das regras da ferramenta ART e os dados da anotação manual do corpus realizada no MMAX. O corpus analisado apresenta um total de 279 descrições definidas, sendo que 131 dessas expressões são classificadas como não anafóricas pela classificação manual, e 94 pela classificação automática, conforme Tabela 1. Para avaliar os ganhos obtidos com as heurísticas propostas, comparamos as medidas de abrangência e precisão das heurísticas com o *baseline* sendo um algoritmo que considera todas as expressões definidas como não anafóricas. A comparação é apresentada na Tabela 2. Com essas heurísticas obtemos 52,6% de abrangência e 73,4% de precisão, o que representa um ganho em relação à precisão obtida com o *baseline*. Considerando-se que apresentamos apenas cinco heurísticas de análise do sintagma pode-se dizer que a abrangência é bastante significativa. Durante o processo de análise dos resultados, erros na classificação foram observados, tais como: algumas descrições definidas sem

complementos (como *"as acusações"*) são não anafóricas, pois fazem parte do título do artigo ("*Citados negam as acusações*"), ou seja, estão na primeira sentença do texto; as descrições definidas constituídas por cláusulas relativas*"*, não estão sendo tratadas pela heurística 3, com por exemplo o pronome relativo "*onde*" em *"o hotel onde se hospeda, em Brasília*, isto se deve ao fato do analisador PALAVRAS considerar o pronome relativo "*onde*" como um advérbio.

Tabela 1. Classificação Manual e Automática

	Não anafóricas	**Anafóricas**	**Total**
Classificação manual	131	148	279
Classificação automática	94	185	279

Tabela 2. Abrangência e Precisão %

	Abrangência	**Precisão**
Baseline	100	46.9
ART + Heurísticas	52.6	73.4

Como trabalhos futuros, pretende-se aumentar o número de características para a identificação das descrições definidas não anafóricas. Essas novas características levariam em conta a posição das descrições definidas no texto, para tratar, por exemplo, as descrições definidas na primeira sentença. Também se consideraria as construções copulares, por exemplo: *"O maior representante do Eurocentres no Brasil é o Stbstudent Travel Bureau"*.

Com base nas heurísticas desenvolvidas, pretende-se além de aumentar o número de características para a identificação das descrições definidas não anafóricas, também utilizá-las em experimentos de resolução de anáforas com uma abordagem de Aprendizado de Máquina Supervisionado com árvores de decisão.

6. Bibliografia

Bean, D. L. and Riloff, E. (1999) "Corpus-based Identification of Non-Anaphoric Noun Phrases". In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, p. 373–380.

Bick, E. (2000) "The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework". PhD thesis, Arhus University, Arhus.

Cardie, C. and Wagstaff, K. (1999) "Noun phrase coreference as clustering". In: Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, p. 82–89.

Gasperin, C.; Vieira, R.; Goulart, R.; Quaresma, P. (2003) "Extrating XML Syntactic Chunks from Portuguese Corpora". In: Traitement Automatique Dês Langues Minoritaires- TALN, Btaz-sur-mer, France.

Gasperin, C., Goulart, R.; Vieira, R. (2003) "Uma Ferramenta para Resolução Automática de Co-referência". Anais do Encontro Nacional de Inteligência Artificial (ENIA 2003), Campinas, SP.

McCarthy, J. F. and Lehnert, G. (1995) “Using decision trees for coreference resolution”. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, p. 1050–1055.

Miorelli, S. (2001) “Extração do Sintagma Nominal em Sentenças em Português”. Dissertação de Mestrado, PUC, Porto Alegre.

Müller, C. and Strube, M. (2000) “MMAX: A tool for the annotation of multi-modal corpora”. In: Proceedings of the IJCAI 2001, Seattle, p. 45–50.

Muller, C.; Stefan, R.; Strube, M. (2002) “Applying Co-training to reference resolution”. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL- 2002), Philadelphia, Penn., p. 352-359.

Ng, V. and Cardie, C. (2002a) “Improving machine learning approaches to coreference resolution”. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Ng, V. and Cardie, C. (2002b) “Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution”. In: Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING-2002), p. 730–736.

Perini, M. (2003) Gramática descritiva do português. São Paulo: Editora Ática, 380 p.

Silva, M. and Koch, I. (1989). Lingüística Aplicada ao Português: Sintaxe. São Paulo: Editora Cortez, 160 p.

Soon, W. M.; Ng, H.wee T.; Lim, D. C. Y. (2001) “A machine learning approach to coreference resolution of noun phrases”. In: Computational Linguistics, p. 521–544.

Uryupina, O. (2003) “High-precision Identification of Discourse New and Unique Noun Phrases”. In: Proceedings of the ACL Student Workshop, Sapporo.

Vieira, R. (1998) “Definite description processing in unrestricted text”. PhD thesis, University of Edinburgh, Edinburgh.

Vieira, V. and Poesio, M. (2000) “An empirically-based system for processing definite descriptions”. In: Computational Linguistics.

Vieira, R.; Salmon-Alt, S.; Schang, E. (2002) “Multilingual corpora annotation for processing definite descriptions”. In: Proceedings of the PorTAL 2002, Faro.

Vieira, R.; Gasperin, C.; Goulart, R.; Salmon-Alt, S. (2003) “From concrete to virtual annotation mark-up language: the case of COMMOn-REFs”. In Proceedings of the (ACL 2003) Workshop on Linguistic Annotation: Getting the Model Right, Sapporo.

Edição de informações sintático-semânticas dos adjetivos na base da rede Wordnet.Br

Ariani Di Felippo, Bento Carlos Dias-da-Silva[1]

[1]Centro de Estudos Lingüísticos e Computacionais da Linguagem (CELiC)
FCL – UNESP – Caixa Postal 174 – Araraquara – SP – Brazil
Núcleo Interinstitucional de Lingüística Computacional (NILC)
ICMC – USP – Caixa Postal 668 – São Carlos – SP – Brazil

{arianidf@uol.com.br,bento@fclar.unesp.br}

Abstract. *This paper proposes the formal inclusion of the "argument structure" and the "subcategorization frame" of adjectives in the Wordnet.Br lexical database. The conclusion outlines such an extension.*

Resumo. *Neste trabalho, propõe-se a inclusão da "estrutura de argumentos" e do "esquema de subcategorização" dos adjetivos na base da rede Wordnet.Br. A conclusão esquematiza essa proposta de extensão.*

1. Introdução

A partir do desenvolvimento da rede WordNet[1] da Universidade de Princeton [Fellbaum 1999], EUA, vários países construíram ou estão construindo suas próprias *wordnets*, dada a importância desse tipo de base lexical na compilação de parcelas de léxicos para o desenvolvimento de diversos sistemas de processamento automático de línguas naturais (PLN). A rede WordNet é, na verdade, uma base relacional, em que unidades lexicais do inglês, pertencentes às categorias dos substantivos, verbos, adjetivos e advérbios, estão organizadas em termos de conjuntos de sinônimos (isto é, os *synsets*) que expressam conceitos lexicalizados. Tais conjuntos relacionam-se entre si em função das cinco relações de sentido: antonímia, hiponímia, meronímia, acarretamento e causa [Vossen 1998]. Além disso, a rede WordNet registra informações periféricas, associadas a cada sentido armazenado. São elas: frases-exemplo e glosas (isto é, definições informais).

A base da rede *wordnet* brasileira (doravante, Wordnet.Br), em desenvolvimento a partir do aplicativo *Thesaurus Eletrônico para o Português do Brasil – TeP* [Dias-da-Silva, 2003], apresenta um total de 17.416 substantivos, 11.078 verbos, 15.073 adjetivos e 1.139 advérbios, estruturados em função das relações de *sinonímia* e *antonímia* [Dias-da-Silva et al. 2002; Dias-da-Silva 2003; Dias-da-Silva e Moraes 2003]. Na fase atual de desenvolvimento da base lexical da Wordnet.Br, estão sendo feitas a análise de

[1] O nome da rede americana é grafado com "N" maiúsculo para diferenciá-la das demais, caracterizando-a, como diz Fellbaum, como "a mãe de todas as Wordnets", construída para essa variante do inglês. Atualmente, várias comunidades de PLN já possuem seus aplicativos no formato *wordnet*. Dentre eles, citam-se as redes originalmente propostas para integrar o núcleo da EuroWordNet: as redes para o inglês britânico, holandês, espanhol, italiano, alemão, sueco, francês, tcheco e estônio. Recentemente, também em fase de construção, cita-se a rede Wordnet.Pr, para o português europeu [Marrafa 2001].

consistência semântica dos synsets e a coleta e seleção das frases-exemplo[2]. Para os pesquisadores do PLN, a base da Wordnet.Br possibilita, por exemplo, a geração de parcelas de léxicos especiais, munidos de conhecimento léxico-semântico, imprescindíveis para o desenvolvimento de diversos sistemas de PLN, tais como: sistemas de tradução automática, de sumarização automática, entre outros [Briscoe e Boguraev 1989], [Saint-Dizier e Viegas 1995], [Dias-da-Silva 1998] e [Palmer 2001]. Ao usuário da língua portuguesa, por sua vez, a base da Wordnet.Br, acoplada a ferramentas computacionais de auxílio à escrita, deverá oferecer a opção de seleção *on line* de palavras sinônimas e antônimas que, por motivos de estilo, precisão, adequação comunicativa, correção ou aprendizagem, o usuário queira substituir [Ilari e Geraldi 1985].

Neste trabalho, em particular, propõe-se a inserção, nessa base, de informações sobre a função de *predicador* do adjetivo do português brasileiro. Nas seções subseqüentes, delineiam-se (i) a motivação (psico)lingüística para a inclusão desse tipo de informação na base da rede Wordnet.Br, (ii) a estrutura atual da base, (iii) o tratamento dado aos adjetivos no projeto Wordnet.Br e (iv) as informações relativas à função de *predicador* que poderão ser especificadas na base. Na conclusão, esquematiza-se a extensão resultante do acréscimo dessas informações.

2. Da motivação (psico)lingüística

Os estudos realizados no domínio da Psicolingüística têm contribuído consideravelmente para a construção de léxicos lingüístico-computacionais [Handke 1995]. A seguir, são feitas considerações a respeito da estrutura global e interna do *léxico mental*[3] com o objetivo de delimitar os subsídios (psico)lingüísticos para a proposta de extensão da base da Wordnet.Br.

2.1. Da Macroestrutura do "léxico mental"

O *léxico mental* (LM) apresenta uma intrincada rede de relações que se estabelecem entre seus constituintes [Mel'čuk 1988]. Essas relações, consideradas *intrínsecas*, são responsáveis pela *macroestrutura* do léxico [Levelt 1993] e as associações estabelecidas entre os itens lexicais distribuem-se em associações (i) paradigmáticas e (ii) sintagmáticas [Biderman 1981].

(1) Das relações paradigmáticas: diz-se que as unidades lexicais pertencem ao mesmo paradigma quando uma puder ser substituída pela outra em um mesmo ponto da cadeia sintagmática; tais relações podem ser: a) morfossemânticas: relações entre os itens que apresentam a mesma raiz (p.ex.: embalar, embalado, embalador; b) léxico-conceituais: relações que se estabelecem entre conceitos lexicalizados: a *hiponímia* (p.ex.: laranjeira é hipônimo de árvore) e a *hiperonímia* (p.ex.: árvore é hiperônimo de laranjeira); c) léxico-semânticas: relações que se estabelecem entre unidades (= formas) lexicais e não entre conceitos; são elas: as relações de *sinonímia* (p.ex.: asfixiado, sufocado) e as de *antonímia* (p.ex.: grande/pequeno).

---

[2] CNPq 09/2001- Processo Nº 552057/01-0.

[3] Entende-se por *léxico mental* a parte do conhecimento lexical do indivíduo delimitada pela sua língua [Bierwisch e Schreuder 1992], [Levelt 1993].

(2) Das relações sintagmáticas: são relações resultantes da combinatória freqüente entre itens lexicais; a principal relação sintagmática da macroestrutura do léxico é a *colocação* (do Inglês: "*collocation*"). O termo *colocação* é aqui entendido como a relação que se verifica entre seqüências de unidades lexicais que co-ocorrem habitualmente como, por exemplo, *feliz aniversário, dias de sol*.

De acordo com o modelo de processamento cognitivo da linguagem de Levelt (1992, 1993), tais relações podem ser consideradas *diretas*, e, conseqüentemente, representadas no interior das entradas do LM. Uma relação léxico-semântica "direta", por exemplo, é aquela em que os sinônimos de um item *x* são listados na entrada lexical de *x*. Essa abordagem, inclusive, é a adotada no modelo *wordnet.*

2.2. Da Microestrutura do "léxico mental"

Além das informações referentes à *macroestrutura* do léxico, as quais podem (ou não) compor as entradas do LM, ressalta-se que a entrada lexical (E) de um item *x* armazena os *lemas* e os *lexemas*, que estão interligados por um ponteiro lexical, isto é, cada *lema* "aponta" para seu *lexema* correspondente (Levelt 1993).

Os **lemas** são representações das propriedades semânticas e sintáticas dos itens lexicais. Observe-se que não se trata, aqui, do sentido em que esse termo é empregado no âmbito da Lexicografia, isto é, a forma canônica de uma unidade lexical. A forma *entrar*, em *João entrou na casa*, pertence à classe V(erbo) e projeta uma estrutura semântica de dois argumentos: [(Agente <*anim*>) (Meta <*loc*>)]; esses argumentos se realizam sintaticamente como um SN (*João*) e um SPrep (*na casa*). A construção de um SV (p.ex.: *entrar na casa*) ou de uma sentença (*O gato entrou na casa*) depende da informação sintática contida no *lema*.

Já os **lexemas** são representações das estruturas morfológica e fonológica das unidades lexicais. Por exemplo, o *lexema* de *entrar* especifica que esse item é formado pelos seguintes segmentos morfológicos: o radical /entr-/, a vogal temática /-a/ e a flexão /-r/; e pelos cinco segmentos fonológicos: /eN/, /t/, /r/, /a/ e /R/. Dessa forma, pode-se dizer que, do ponto de vista (psico)lingüístico, as informações semânticas armazenadas na base da rede Wordnet.Br são referentes às relações que se estabelecem entre os itens do LM. Essa rede equaciona, do ponto de vista computacional, parte das relações semânticas responsáveis pela estrutura global ou *macroestrutura* do LM.

A extensão dessa base aqui proposta consiste na inserção da informação sintático-semântica de *predicador* que parte dos adjetivos da língua desempenha. Nesse sentido, além de abrigar as relações léxico-semânticas que se instauram entre os adjetivos, ela deverá também abrigar as informações sobre o *lema* de cada adjetivo predicador, ou seja, informações responsáveis pela *microestrutura* do LM [Bierwisch e Schreuder 1992], [Handke 1995].

3. Da estrutura atual da base da Wordnet.Br

Como mencionado, a elaboração da base da Wordnet.Br teve como ponto de partida o TeP, que foi elaborado segundo os princípios da rede americana. Da metodologia proposta por Miller e Fellbaum (1991) para a construção da rede WordNet, foram utilizadas três noções básicas no desenvolvimento do Tep [Dias-da-Silva et al 2002]: (i) o *método diferencial*, que pressupõe o princípio de ativação de conceitos por meio de

um conjunto de formas lexicais relacionadas pela sinonímia, o que elimina a necessidade de especificação do valor semântico para o sentido de uma entrada lexical; (ii) a noção constitutiva de *synset*, isto é, o conjunto de sinônimos; (iii) a noção de *matriz lexical*, que postula uma correspondência biunívoca entre sentido e *synset*.

Com essa metodologia, a relação de sinonímia passa a ser representada formalmente pela relação lógica de pertença (x é sinônimo de y $\leftrightarrow$ x $\wedge$ y $\in$ A, em que A é um *synset*). A antonímia, por sua vez, é representada por uma relação entre conjuntos (x é antônimo de y $\leftrightarrow$ x $\in$ A e y $\in$ B, A e B são *synsets* e A e B estão relacionados pela relação de antonímia). O Esquema 1 ilustra a estrutura da base do TeP e, conseqüentemente, da Wordent.Br:

(1) Unidade de Entrada n (categoria sintática *x*)
- Sentido n1
  - Synset
  - Synset de antônimos
- Sentido n.m

Nesse esquema, "n" é o número de identificação da unidade de entrada, "x" é uma variável que representa uma das quatro categorias gramaticais (substantivo, verbo, adjetivo ou advérbio), "n.1...n.m" são números que identificam cada sentido da unidade de entrada n.

4. Do tratamento dado aos adjetivos na Wordnet.Br

Para a compilação dos conjuntos de sinônimos e antônimos de adjetivos do TeP, partiu-se do princípio de que os adjetivos do português, assim como os do inglês e do espanhol, podem ser divididos em duas classes: os *qualificadores* (QLs) e os *classificadores* (CLs) [Quirk et al 1991], [Borba 1996], [Demonte 1999] e [Neves 2000].

Os **qualificadores** indicam o valor de uma propriedade ou atributo do substantivo com o qual se liga. Dessa forma, dizer "X QL" ou "X é QL" pressupõe um atributo A, tal que A(x) = QL. Por exemplo, dizer "torre *alta*" ou "a torre é *alta*", pressupõe um atributo ALTURA, tal que ALTURA(torre) = *alta*. Além disso, pressupõe-se que: (i) os atributos são *bipolares*, isto é, os adjetivos *alto* e *baixo* são antônimos e expressam os valores dos pólos do atributo ALTURA; (ii) atributos podem ser *graduáveis* (contínuos) ou *não-graduáveis* (dicotômicos), por exemplo: o atributo graduável ALTURA varia em um contínuo de "alturas" entre os valores polares "alto" e "baixo", isto é, os valores dos extremos do atributo ALTURA; já o atributo SEXO, não-graduável, apresenta apenas dois valores: *macho* e *fêmea*; e, por isso, são denominados *dicotômicos*. Dentre as principais características dos QLs estão [Casteleiro 1981], [Gross, Fischer e Miller 1995]: (i) a nominalização da propriedade que expressam (p.ex.: muro *alto* → a *altura* do muro); (ii) a gradação (p.ex.: o muro é muito *alto*); (iii) a comparação (p.ex.: a torre é mais *alta* do que a pirâmide).

Já os **classificadores** colocam a denotação do substantivo com o qual ocorrem numa subclasse, nomeando-a, p.ex.: o adjetivo classificador *cambial* no SN *a reforma cambial*. Observe-se que a paráfrase *do câmbio* sinaliza que *cambial* liga a entidade "reforma" a outra, exterior a ela: o "câmbio" [Borba 1996], [Neves 2000], [Basílio e

Gamarski 2002]. Enquanto os QLs possuem as propriedades descritas em (i), (ii) e (iii), o mesmo não ocorre com os Cls (p.ex.: *a cambiabilidade da reforma/ *a reforma é muito cambial/ *a reforma é mais cambial do que a crise.)[4]. Enquanto os adjetivos QLs expressam "qualidades" ou "valores de atributos" dos substantivos, os CLs são comumente definidos, em obras lexicográficas, por meio de paráfrases como "de ou pertencente/ relativo a X" [Miller e Fellbaum 1991].

5. Da sistematização das "novas" informações sobre os adjetivos

As duas funções sintático-semânticas básicas dos adjetivos relacionam-se com sua posição: **posição adnominal** (Padn) e **posição predicativa** (Ppred). Tanto os QLs quanto os CLs podem ocorrer em Padn; a Ppred, no entanto, não é exclusiva, mas sim característica da subclasse dos QLs. Isso porque há certos CLs que admitem a função predicativa quando em condições contextuais específicas [Casteleiro 1981]: (i) com construções contrastivas (p.ex.: Estas viaturas são *municipais*; aquelas, não); (ii) com repetição do núcleo do sintagma nominal (p.ex.: Esta estrada é uma estrada *vicinal*). Os CLs, que comumente ocorrem apenas em Padn, são, na verdade, complementos dos substantivos (p.ex.: câmara municipal > câmara do município) com os quais ocorrem ou têm valores adverbiais (p.ex.: matador profissional).

Focalizando os QLs, ressalta-se que esses, quando em Padn, expressam o valor de atributo *preexistente* ao julgamento do falante (p.ex.: rapaz *pobre*). De acordo com essa visão, pressupõe-se que a categoria do substantivo seja um conjunto de propriedades ou atributos e que a função do adjetivo QL (em Padn, ou em função modificadora) é a de preencher o valor de um desses atributos. Os QLs, quando em Ppred, instauram verdadeiro processo de *predicação* (p.ex.: Aquele rapaz é *pobre*).

5.1. Os adjetivos *predicadores* ou *valenciais*

Sendo, portanto, **predicador** (PR), o QL designa um "estado-de-coisas", isto é, algo que pode ocorrer em algum mundo (real ou mental) (Dik, 1997). No interior da predicação, estabelecem-se as propriedades ou relações especificadas pelo PR. A *valência*, então, pode ser entendida como a relação abstrata do PR com os argumentos (As) que dele dependem [Neves 1996].

O termo *valência* pode ser usado em três níveis: *valência lógico-semântica, valência sintática* e *valência semântica*.

a) Quanto à valência lógico-semântica dos adjetivos

Sendo o nível mais abstrato, diz respeito ao número de As que um PR pode ter. Há duas interpretações possíveis para a valência adjetival; na primeira, consideram-se apenas os constituintes diretamente dependentes dos adjetivos; na segunda, considera-se, como argumento adjetival, o constituinte em função de sujeito [Busse e Vilela 1986]. Neste trabalho, optou-se pela segunda interpretação. Dessa forma, considera-se que os adjetivos PRs ou valenciais do português podem ser de quatro tipos [Borba 1996], como descrito na Tabela 1.

[4] O símbolo "*" indica agramaticalidade.

Tabela 1. Valência lógico-semântica: tipologia

Tipologia	Descrição e exemplificação
Valência 1 (V_1)	Projeta um argumento lógico-semântico. P.ex.: Meu pai (A1) era *alto*, loiro e de olhos azuis.
Valência 2 (V_2)	Projeta dois argumentos lógico-semânticos. P.ex.: Renunciou à convicção porque ela (A1) não era *útil* a seus propósitos (A2).
Valência 3 (V_3)	Projeta três argumentos lógico-semânticos. P.ex.: O réu (A2) era *condenável* à morte (A3) pelo juiz. (A1).
Valência 4 (V_4)	Projeta quatro argumentos lógico-semânticos. P.ex.: A carga (A2) era *transportável* do estaleiro (A3) para o navio (A4) por guindastes (A1).

b) Quanto à valência sintática dos adjetivos

Esse nível trata da função sintática e/ou da categoria sintagmática (e/ou morfossintática) dos As realizados na sintaxe, p.ex.: em *Aquele rapaz é pobre*, o adjetivo *pobre* requer (ou projeta) um argumento lógico-semântico (A1), que, sintaticamente, realiza-se sob a forma do sintagma nominal sujeito ("aquele rapaz"). A valência sintática também pode ser entendida como **esquema de subcategorização** [Raposo 1992]. Salienta-se que nem todos os As lógico-semânticos são realizados na sintaxe. No caso dos adjetivos valenciais (ou seja, aqueles em Ppred), observa-se que eles partem sempre de um índice *1*, o sujeito. Os demais argumentos são opcionais.

c) Quanto à valência semântica dos adjetivos

Esse nível trata das relações semânticas que se estabelecem entre o PR e os As. A valência semântica é também designada **estrutura de argumentos** [Grimshaw 1992]. Mais especificamente, são observadas, nesse nível, as funções temáticas (= papéis) dos As e as restrições selecionais que o PR impõe sobre seus argumentos. Por exemplo: em *Paulo era descendente de italianos*, o adjetivo *descendente* projeta dois argumentos lógico-semânticos, que se realizam sob a forma de SN ("Paulo") e de SPrep ("de italianos") e que, do ponto de vista semântico, recebem papel temático *Objetivo* ("Paulo") e *Origem* ("de italianos"), sendo que ambos precisam ser do tipo semântico <humano> [Borba 1996].

6. Conclusão

A partir da breve exposição sobre os adjetivos valenciais do português, propõe-se a ampliação da estrutura original da base da Wordnet.Br para que as seguintes informações possam ser inseridas: (i) a valência lógico-semântica; (ii) a valência sintática ou esquema de subcategorização; (ii) a estrutura de argumentos ou valência semântica.

6.1 Da extensão da informação-base da Wordnet.Br

Em suma, a classificação esboçada na seção anterior permite estender a informação original (Esquema 1) relativa aos QLs da Wordnet.Br com os seguintes tipos: QL = {PR} = {Adj_V1} + {Adj_V2} + {Adj_V3} + {Adj_V4}. Assim, paralelamente ao

Esquema 1, propõe-se que aos QL da base da Wordnet.Br estejam associadas informações sobre a *valência lógico-semântica* e sobre a *estrutura de argumentos* ou *valência semântica*. O Esquema 2 exemplifica essa extensão.

(2) Unidade de Entrada n (Tipo semântico QL)
- Função sintático-semântica PR
  - Valência n.1 <*papéis temáticos dos As + restrições de seleção*>
  - Valência n.4

Nesse esquema, "n" é o número de identificação da unidade de entrada; "QL" indica que o adjetivo é do tipo "qualificador"; "PR" indica que a função sintático-semântica é a de "predicador"; "n.1...n.4" indicam qual o subtipo lógico-semântico, (ou seja, V1, V2, V3 e V4), ao qual é associada a *valência semântica* (papéis temáticos + restrições selecionais) do adjetivo propriamente dito.

Salienta-se que indexações apropriadas deverão permitir, quando pertinente, o relacionamento entre as entradas estruturadas em termos das relações léxico-semânticas (sinônima e antonímia) e as entradas estruturadas em termos das informações sintático-semânticas aqui propostas. No caso específico do Esquema 2, indexações deverão permitir o relacionamento entre *valência* e *sentido*, este armazenado na base da Wordnet.Br em função dos conjuntos de sinônimos. O Esquema 3 ilustra essa possível indexação.

(3)
- Adj **desgostoso**
  - Sentido1
    - Synset

⟷

- QL **desgosotoso**
  - PR
    - V2 <[Ob:anim][$_{com}$[Ca:inanim]]>

No exemplo em (3), observa-se que a valência semântica (ou estrutura de argumentos) descrita entre os símbolos "< >" está associada ao Sentido1 do adjetivo *desgostoso*. Ressalta-se que para a especificação dessa valência, tomar-se-á como base a frase-exemplo. Ao adjetivo *desgostoso*, no Sentido1, por exemplo, está associada a frase "O príncipe Charles está desgostoso e escandalizado com as acusações contra sua mulher, a princesa Diana". A partir da frase-exemplo, identificam-se as relações semânticas que o PR estabelece com os argumentos e as restrições que ele impõem sobre os mesmos. Nessa frase, o adjetivo PR estabelece uma relação de "causa" com o argumento que se realiza na forma do SPrep "com as acusações" e, por isso, esse A recebe o papel temático Ca(usativo); já o A que se realiza na forma do SN sujeito "o príncipe Charles" recebe papel temático Ob(jetivo). Dessas observações, elaborar-se-á uma estrutura do tipo <[Ob:*anim*][$_{com}$[Ca:*inanim*]]>.

Uma vez especificada a estrutura de argumentos que "traduz" o(s) sentido(s) de um adjetivo QL, poder-se-á generalizá-la para os demais membros do *synset*. Assim, os adjetivos {anojado; desagradado; descontente; desgostoso; dissaborido; dissaboroso; malcontente; penalizado; triste} poderão herdar a valência especificada no exemplo em (3).

6.2 Da extensão da informação periférica da Wordnet.Br

Se, por um lado, as valências lógico-semântica e semântica poderão ser indexadas à estrutura original da base da Wordnet.Br (Esquema 1), por outro lado, a valência sintática, ou, como já se disse, *esquema de subcategorização*, poderá ser associada a uma das informações classificadas aqui como periféricas, no caso, a frase-exemplo. Diz-se periférica pelo fato de que essas frases, assim como as *glosas*, não constituem os *synsets*. Merece destaque, entretanto, o fato da frase-exemplo fornecer o importante contexto de uso do adjetivo. Assim, na extensão que aqui se propõe, o *esquema de subcategorização*, ao estar associado à frase-exemplo, constitui o que poderia ser considerado o "comentário sintático" desta. Em (3), por exemplo, o esquema de subcategorização <[SN] [SPrep (com SN)]]> seria associado à frase-exemplo "O príncipe Charles está desgostoso e escandalizado com as acusações contra sua mulher, a princesa Diana".

Assim, com esse acréscimo, a manipulação da base da rede Wordnet.Br poderá gerar listas de formas, para a compilação de léxicos monolíngües, que, além de fornecerem as relações (léxico-semânticas) que se instauram entre os adjetivos, fornecerão também sua valência e seu esquema de subcategorização, isto é, informações sobre o *lema* do adjetivo, ou seja, sobre propriedades sintático-semânticas responsáveis pela *microestrutura* do léxico da língua [Bierwisch e Schreuder 1992], [Handke 1995].

Referências

Basílio, M. e Gamarski, L. "Adjetivos denominais no português falado". In: Castilho, A. T. de (org.). Gramática do Português Falado – v. IV. 2ª ed. Campinas, UNICAMP, p. 629-650, 2002.

Bierwisch, M. e Schreuder, R. (1992) "From concepts to lexical items". Cognition, 42, p.23-60.

Biderman, M. T. C. "A estrutural mental do léxico". In: Estudos de Filologia e Lingüística - Homenagem a Isaac Nicolau Salum. São Paulo, Editora da USP; T. A. Queiroz, p. 131-45, 1981.

Borba, F. S. "Uma gramática de valências para o português". São Paulo, Editora Ática, 1996.

Busse, W. e Vilela, M. "Gramática de valências". Coimbra, Almedina, 1986.

Briscoe, E. J. e Boguraev, B. (eds) "Computational lexicography for natural language processing". London/New York, Longman, 1989.

Casteleiro, J. M. "A sintaxe transformacional do adjetivo". Lisboa, Instituto Nacional de Investigação Científica, 1981.

Demonte, V. (1999) "Semántica composicional y gramatica: los adjetivos en la interficie léxico-sintaxis". Revista Española de Lingüística, 29, v.2, p. 283-316.

Dias-da-Silva, B. C. (1998) "Bridging the gap between linguistic theory and natural language processing". In: Proceedings of the 16th international congress of linguistics. Oxford, Elsevier Sciences, n. 16, p. 1-10.

Dias-da-Silva, B. C. (2003) "O TeP: construção de um thesaurus eletrônico para o português do Brasil". Boletim da Associção Brasileira de Lingüística (ABRALIN).

Fortaleza: Imprensa Universitária, v.26, número especial, p.86 - 89.

Dias-da-Silva, B. C. e Moraes, H. R. (2003) "A construção de thesaurus eletrônico para o português do Brasil". Alfa , v.47, n.2, p.101 - 115.

Dias-da-Silva, B. C., Oliveira, M. F. e Moraes, H. R. (2002) "Groundwork for the development of the Brazilian Portuguese Wordnet". Advances in natural language processing. Berlin, Springer-Verlag, p.189-196.

Dik, S. C. "The theory of functional grammar". Berlin, New York: Mouton de Gruyter, 1997.

Fellbaum, C. (ed.) "Wordnet: an electronic lexical database". Cambridge, The MIT Press, 1999.

Grimshaw, J. "Argument structure". Cambridge, The MIT Press, 1992.

Gross, D., Fischer, U. e Miller, A. (1989) "The organization of the adjectival meaning". Journal of Memory and Language, 28, p. 92-106.

Handke, J. "The structure of the Lexicon: human versus machine". Berlin, Mouton de Gruyter, 1995.

Ilari, R. e Geraldi, J. W. "Semântica". São Paulo, Editora Ática, 1985.

Levelt, W. J. M. (1992) "Accessing words in speech production: stages, processes and representations". Cognition, 42, p.1-22.

____. "Speaking: to intention to articulation". Cambridge, The MIT Press, 1993.

Lyons, J. "Language and linguistics. An introduction". Cambridge, CUP, 1981.

Marrafa, P. (2001) "WordNet do Português – Uma base de dados de conhecimento lingüístico". Lisboa: Instituto Camões, 2001.

Mel'čuk, I. "Dependency Syntax: theory and practice". The SUNY Press, Albany, N.Y, 428p. 1988.

Miller, G. A. e Fellbaum, C. (1991) "Semantic networks of English". Cognition, v.41, n.1-3, p.197-229.

Neves, M. H. M. "Gramática de usos do português". São Paulo, Editora UNESP, 2000.

____. "Estudo da estrutura argumental dos nomes". In: Kato, M. A. (org.) *Gramática do Português Falado* V: Convergências. Campinas: Ed. Unicamp/FAPESP, p. 119-154, 1996.

Palmer, M. (2001) "Miltilingual resources – Chapter 1". In: Hovy, E., et al. (eds.). Linguistica Computazionale, v.14-15.

Quirk, R. et al. "A Comprehensive Grammar of the English Language". London, Longman, 1991.

Raposo, E. P. "Teoria da gramática: a faculdade da linguagem". Lisboa, Caminho, 1992.

Saint-Dizier, P. e Viega, E. "Computational lexical semantics". Cambridge, Cambridge University Press, 1995.

Vossen, P. et al. (1998). "The EuroWorNet base concepts and top ontology". <http://www.illc.uva.nl/EuroWordNet/docs.html>. Fev. 2003.

Locution or collocation: comparing linguistic and statistical methods for recognising complex prepositions

Claudia Oliveira[1], **Cícero Nogueira**[1], **Milena Garrão**[2]

[1]Departamento de Engenharia de Sistemas
Instituto Militar de Engenharia, Rio de Janeiro

[2]Departamento de Letras
Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro

{cmaria, nogueira}@de9.ime.eb.br, migarrao@uol.com.br

Abstract. *Multi–Word Expressions (MWE) include a large range of linguistic phenomena, such as nominal compounds and institutionalized phrases. These expressions, which can be syntactically and/or semantically idiosyncratic in nature, are frequently used in everyday language as well as in formal contexts. In this work we investigate a type of MWE, the complex preposition (CP). The purpose is to establish the relationship between the notions of locution - the linguistic view of CPs - and collocation - the statistical view - when we look into the corpus, and to consider how these notions can be applied to the delimitation of CPs.*

Resumo. *Expressões Multivocabulares (EMV) englobam uma grande coleção de fenômenos lingüísticos, tais como compostos nominais e frases institucionalizadas. Essas expressões, que podem ser sintaticamente e/ou semanticamente idiossincráticas, são freqüentes na linguagem oral assim como em contextos formais. Neste trabalho nós investigamos um tipo de EMV, a locução prepositiva (CP). O objetivo é estabelecer o relacionamento entre as noções de locução - a visão lingüística de CPs - e colocação - a visão estatística - quando analisamos o corpus, e consideramos como essas noções podem ser usadas na delimitação de CPs.*

1. Introduction

In recent years, there has been a growing interest in the issues involved in dealing with Multi–Word Expressions (MWEs) in most areas of Natural Language Processing (NLP). MWEs include a large range of linguistic phenomena, such as nominal compounds (e.g. “table cloth”), and institutionalized phrases (e.g. “fish and chips”). These expressions, which can be syntactically and/or semantically idiosyncratic in nature, are very frequent in everyday language as well as in formal contexts. Applications that require some degree of semantic interpretation (e.g. machine translation, question-answering, summarisation, generation) and require tasks such as parsing and word sense disambiguation are particularly sensitive to MWEs’ delimitation problems.

Brazilian grammarian Mattoso Câmara Jr [Mattoso Câmara Jr, 1984] considers a locution to be “a conjunction of two words which maintain their phonetic and morphemic

individuality, but make up a signifying unit for a specific function". The locutional character of an expression relies on the fact that it is a signifying block with a given role and with a distinguished part of speech. He emphasizes that the nouns forming the prepositional locution have already undergone grammaticalisation; in other words, they have been through a "process that consists of turning simple, lexical semantically full words into grammatical words". What should be considered, therefore, is the signifying block and not the meaning of each of the items belonging to the locution. Even though it is not explicitly discriminated in [Mattoso Câmara Jr, 1984], we notice that locution is mostly used with respect to functional classes such as adverbial, conjunctional or prepositional locutions. In a descriptive grammar one would find the following definition for an adverbial locution: "... two or more words working as an adverb".

Collocations, on the other hand, are defined in terms of frequency of co–occurrence. Manning and Schütze [Manning and Schütze, 1999] define collocations as "two or more words that correspond to some conventional way of saying things", which highlights the habitual place of a word in relation to another.

There are some important common points between the notions of locution and collocation. First of all, both present limited compositionality of meaning. This is the most distinguishable feature of MWEs. Another characteristic of MWEs present in both notions are non–substitutability of its components by near synonyms and non–modifiability of the phrase as a whole, for instance the impossibility of insertion of other lexical items.

The purpose of this work is to establish the relationship between the notions of locution and collocation when we look into the corpus, and to consider how they can be applied to the delimitation of Complex Prepositions (CPs). There are some recent works with similar aims and different target languages. In [Trawinski, 2003], the focus is on the representation of the syntax of German CPs in HPSG. More closely related to our work is [Bouma and Villada, 2002], in which a list of Dutch collocations of the form "Prep_1 NP Prep_2" is extracted from a corpus and analysed by human judges to determine which ones are CPs and therefore to establish the effectiveness of statistical methods.

The remainder of this paper is organised as follows. In section 2 we review the grammatical notion of CPs and a summary of the operational criteria for the delimitation of CPs. In section 3 we describe the statistical approaches to extracting collocations from a corpus. In section 4 we describe the data and the statistical experiments that we carried out, comparing the results with a list of well established CPs. Final comments and conclusions are presented in section 5.

2. Complex prepositions: linguistic facts

A Complex Preposition is a type of MWE that functions as one preposition. In Portuguese and other romance languages, a CP can have a variety of internal structures, such as: "Adverb Prep" (**dentro de**), "Prep Prep" (**por sobre**), "Prep Prep Prep" (**por trás de**), "Prep V Prep (**a partir de**) and "Prep N Prep" (**de acordo com**). We address the delimitation of the latter type of CPs for two reasons. First, because they are more numerous and more frequent. Secondly, given the utmost importance of spotting noun phrases (NP) in text systems, parsing prepositional structures such as "Prep_1 N Prep_2 X" prevents the fragment "N Prep_2 X" from being detected as a noun phrase, i.e. the prepositional structure

is a negative pattern to be used in the extraction of noun phrases from texts.

Prescriptive grammarians of the Portuguese language have not treated the concept of CPs systematically. They resort to using lists to describe the phenomenon that is not restricted to such a simple formal representation. The universally accepted list of simple prepositions, also called the list of essential prepositions, is easy to characterise, because it is a finite set "a, até, após, contra, para, por, de, desde, ante, perante, trás, sob, sobre, com, em, entre, sem". Listing CPs, however, seems to generate at least two immediate problems that are clearly related. One, of a practical nature, reveals the discrepancies between the listings of different grammarians. The other, of a more theoretical nature, confirms the position that CPs constitute an open class.

The grammarians' definitions agree upon two main aspects. Under the formal aspect, they all present a preposition as the last element of what is considered a CP. Under a functional aspect, they claim that the CP is applied as a preposition. One may say, however, that the definitions do not exhaustively describe the phenomenon because they fail to provide consensual and operational criteria to identify the CP.

From a functional perspective, Dias [Dias, 2002] comes to a conclusion that the CPs are a subgroup of prepositions, since they present more similarities than differences when compared to simple prepositions. She considers a CP to be an unfolded preposition, carrying the same syntactic role (i.e. heading prepositional phrases) and the same discourse function (i.e. connectors). The fact that Quirk et al. [Quirk et al., 1978] adopts the terms simple and complex prepositions for the English language confirms the level of generalization regarding this functional perspective.

One focus of our investigation is the formulation of systematic criteria for recognizing CPs. Considering that the class of CPs is open and productive in Portuguese, the task of characterising it goes beyond the trivial enumeration of expressions. It is important to keep in mind that the resulting criteria is to be employed in spotting multi–word prepositions in a sequence of words that includes a noun, in order to rule out the detection of noun phrases containing that noun. In other words, we are interested in distinguishing sequences introduced by a CP followed by a NP, with the structure i. *CP(Prep*$_1$ *N*$_1$ *Prep*$_2$*) NP* from prepositional phrases with the structure ii. *Prep*$_1$ *NP(N*$_1$ *Prep*$_2$ *NP)*.

The order in which the criteria are presented is not incidental, but rather it reflects the decisiveness of the corresponding testing mechanism in spotting the CP. On the other hand, it is not the case that a single test, or even the combination of all the tests, will result in a foolproof interpretation of the expression. We apply them to gather evidence in favour of structure i. or ii. At the same time, it is possible that a CP will test positive to some criteria and not to others. In summary, the tests or any groups of tests, are neither necessary nor sufficient to categorically determine the interpretation of a given expression.

Criterion 0: A priori lexicalisation The most decisive test for frozen CPs is the inexistence of the noun in any other context in the language, for example **em prol de** and **em cima de**. According to this criterion, the CP can be unambiguously recognised as a frozen expression, precluding the need for further testing. It precedes any other criterion for it is the most decisive and the cheapest, from a computational standpoint, since it does not require syntactic parsing.

Criterion 1: Substitution This criterion is based on the notion presented by some grammarians that a CP can normally be substituted by a simple preposition or by another CP. For instance, the sequence [**em virtude de**] ("in virtue of") in 2.1 can be replaced by the CP [**por causa de**] ("because"). In example 2.2, the sequence [**em companhia de**] ("in the company of") can be replaced by the preposition [**com**] ("with").

2.1 *Senna morreu* **em virtude de** *uma falha mecânica da Williams ... (por causa de)*

2.2 *Ele vai passar os próximos meses* **em companhia de** *outros dois cosmonautas que chegaram à Mir no mês passado. (com)*

This test is very attractive at first glance, but presents a few problems when it comes to its implementation. We encountered several examples in the corpus for which we could not find a suitable substitution, such as in 2.3.

2.3 *Só é legal a doação feita* **em troca de** *bônus no valor correspondente.*

Criterion 2: Valency of the preceding verb This criterion uses a very important feature of the syntactic structure of the sentence. If the preposition $Prep_1$ in a sequence "V $Prep_1$ N $Prep_2$ X" is part of the valency of the verb V then "$Prep_1$ N $Prep_2$ X" is the complement in the verb phrase and consequently "$Prep_1$ N $Prep_2$" is not a CP. The following examples make this statement clearer. In 2.4, the preposition **em** ("in") is the head of an essential complement of the verb **aplicar** ("to apply") therefore the sequence "em processo de execução" is interpreted as "PP(em SN(em processo de execução))".

2.4 *Esta multa só se [aplica em] processo de execução; não cabe em procedimento de jurisdição voluntária (JTJ 151/90).*

On the other hand, in example 2.5, a similar analysis is not possible. In this case, the sequence "em processo de" is clearly a CP, which shows semantic non-compositionality and which can be substituted by the simple preposition **em** ("in") (criterion 1).

2.5 *Algumas empresas [***em processo de***] finalização de balanço anual estõ remetendo recursos para o exterior para fugir de obrigações fiscais.*

Criterion 3: Insertion of a determinant This criterion consists of analysing the consequences of inserting a determinant into the sequence "$Prep_1$ N $Prep_2$" to obtain "$Prep_1$ Det N $Prep_2$". The idea is to break the integrity of the expression, so that a CP, as a unit, should not allow such insertion. Sometimes the insertion is simply impossible, as in the case of the definite article **o**, in the contraction [**em + o = no**], in example 2.6.

2.6 *Dizem que o Sr. articula a saída de Bisol* **em favor de***/*no favor de Roberto Freire.*

In other cases we verified a significant semantic change in the resulting expression, as in the case of the definite article **a**, in the contraction [**em + a = na**], in example 2.7.

2.7 *A maioria dos reféns foi libertada na quinta-feira* **em troca de/ na troca de** *armas e drogas.*

There are also cases in which the semantic impact of the determinant is very slight, as in 2.8, which leads us to recommend the use of this criterion to corroborate others, rather than to be used on its own.

2.8 *Os restos de folículos produzem a progesterona, hormônio que, juntamente com os estrógenos, prepara a parede do útero para receber o embrião* **em caso de/ no caso de** *gravidez.*

A variation of the insertion of a determinant is the inflexion of the noun in the sequence "Prep_1 N Prep_2". In example 2.9 such transformation would entail not only the plural, but also a change in the meaning and the parsing of the sentence, as **em forma de** is recognised as a CP. On the other hand, example 2.10 shows a simple plural inflexion of **em forma de**.

2.9 *Estes elementos existem apenas em forma de átomos separados.*

2.10 *Cerca de 57% de todas as florestas tropicais, ambientes mais diversos em formas de vida em todo o mundo, estão representados na região Neotropical.*

This criterion could be generalised to cover the insertion of any lexical material such as pronouns or adjectives.

3. Collocations: statistical tests

The most straightforward method for finding collocations in a corpus is computing the frequency of word pairings (bigrams). Frequent bigrams have a good chance of being collocations. The problem with this approach is that the most frequent words of the language will tend to combine more. In our particular case, the part–of–speech (POS) pattern of the expressions we are investigating contains at least two function words: the prepositions. They are so frequent that their combination with frequent nouns will be wrongly analysed as collocations. The statistical tests we selected balance the effect of individual frequencies, and measure whether a sequence of words occurs more often than would be expected on the basis of individual word frequencies. Therefore, these tests are often used to determine whether two co-occurring words are potential collocations.

Log-likelihood score. A typical problem of statistics is determining whether something is a chance event or linked to another. We want to know whether a bigram is a collocation or whether it appears together by chance. This type of hypothesis testing requires two hypotheses to be formulated (see [Manning and Schütze, 1999]):

$$H_1 : P(w_1|w_2) = P(w_1|\neg w_2)$$
$$H_2 : P(w_1|w_2) \neq P(w_1|\neg w_2)$$

H_1 assumes that the two words are independent and H_2 assumes otherwise. H_2 is the collocation hypothesis and the log-likelihood test measures how much more likely H_2 is than H_1.

Pearson's χ^2 test. The χ^2 test computes the observed frequencies of the following bigrams: w_1w_2, $\neg w_1w_2$, $w_1\neg w_2$ and $\neg w_1\neg w_2$. If the differences between observed frequencies and expected frequencies for independence are large, then the bigram w_1w_2 is a possible collocation. This computation is done by (see [Manning and Schütze, 1999]):

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $i \in \{w_1, \neg w_1\}$, $j \in \{w_2, \neg w_2\}$, O_{ij} is the observed frequency of bigram ij and E_{ij} is the expected frequency.

Mutual Information. This test compares the probability of observing two words $w_1 w_2$ together with the probabilities of observing them independently in a given corpus, computed by (see [Manning and Schütze, 1999]):

$$I(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

The collocation tests we selected take bigrams as inputs but the expressions we are looking at consist of 3 or 4 words. In order to apply the bigram tests to our data-set, we transformed the $w_1\ w_2\ w_3\ w_4$ strings into $w_1\ w_2\_w_3\_w_4$ or $w_1\_w_2\_w_3\ w_4$ strings. In our pattern "Prep$_1$ Det N Prep$_2$", we assume that "Det N Prep$_2$" can be seen as a unit or alternatively that "Prep$_1$ Det N" can be seen as a unit.

4. Working with the corpus

We used the corpus CETENFolha (Corpus de Extractos de Textos Eletrônicos NILC/Folha de São Paulo), containing around 24 million words in Brazilian Portuguese, built by the project Computational Processing of Portuguese from the texts of Folha de S. Paulo belonging to the corpus NILC/São Carlos, compiled by Núcleo Interinstitucional de Lingüística Computacional (NILC) [CETENFolha, 2004].

Muitas	[muito] <quant> DET F P @SUBJ>
de	[de] <sam-> PRP @N<
as	[o] <-sam> <artd> DET F P @>N
prioridades	[prioridade] N F P @P<
de	[de] <sam-> PRP @N<
o	[o] <-sam> <artd> DET M S @>N
novo	[novo] ADJ M S @>N
governo	[governo] N M S @P<
coincidem	[coincidir] <fmc> V PR 3P IND VFIN @FMV
com	[com] PRP @<PIV
as	[o] <artd> DET F P @>N
prioridades	[prioridade] N F P @P<
de	[de] <sam-> PRP @N<
o	[o] <-sam> <artd> DET M S @>N
PT	[PT] PROP M S @P<

Table 1: Extract from the corpus CETENFolha

4.1. Extracting the pattern

We extracted the instances of the pattern "PRP (DET)? N PRP" (coded in the CETENFolha tag-set) where "DET" is optional, and arranged the resulting expressions in the format required by the statistical package NSP [Banerjee and Pedersen, 2003].

em_PRP os_DET tempos_N em_PRP
de_PRP as_DET prioridades_N de_PRP
com_PRP as_DET prioridades_N de_PRP
para_PRP a_DET realização_N de_PRP
sob_PRP o_DET comando_N de_PRP
em_PRP a_DET volta_N de_PRP
de_PRP lista_N em_PRP
para_PRP o_DET sucesso_N de_PRP
de_PRP os_DET recursos_N para_PRP
de_PRP o_DET orçamento_N de_PRP

Table 2: The first 10 candidate expressions extracted from the corpus

In the corpus, each line contains a word followed by its morphosyntatic attributes. We extracted sequences of words with POS tags "PRP N PRP" or "PRP DET N PRP". Table 1 shows an extract of the corpus CETENFolha, with the target pattern highlighted.

The extracting program takes as input the corpus and a list of stopwords. This list contains nouns which are not to be allowed as part of CPs, such as months, days of the week and longhand numbers. Upper case nouns and nouns containing numerical digits are also eliminated, as these involve names, acronyms, dates, numbers, etc. which should not be part of potential CPs. Table 2 shows the first 10 candidate expressions extracted from the corpus into a file.

We found a total number of 632,005 matching strings, with 134,803 distinct ones. In the generated file the words are followed by their POS tags in order to enable bigram formation by the package NSP. The bigrams are of the form (PRP_(DET)?_N, PRP) and (PRP, (DET)?_N_PRP).

4.2. Analysing results

For each one of the three tests we selected - Log-likelihood score (LL), Pearson's χ^2 (χ^2) and Mutual Information (MI) tests - and each of the two forms of bigrams, the output of the statistical package is a list of candidate CPs and their ranking according to the test.

Table 3 shows the top–ranked expressions, according to raw frequencies and to tests LL and MI. The data suggests that we are likely to find CPs amongst frequent "PRP (Det)? N PRP" strings. The results from the statistical tests filter the patterns with strong collocational properties, but the improvement is not noticed in the top 20 strings, considering LL and MI. The performance of χ^2 was very poor in this interval, therefore it has not been included in the table.

Given the amount of variation within the class of CPs and the amount of disagreement between linguists and grammarians, one cannot expect to find an exhaustive listing of CPs. This leaves us with the problem of how to validate the statistical results. The best solution is to present a list of statistically discovered collocations to a group of lexicographers, and let them select the CPs, but this is a costly enterprise, which we have not managed to accomplish yet.

raw frequency	LL: PRP_N PRP	LL: PRP N_PRP	MI: PRP_N PRP	MI: PRP N_PRP
em relaç ˜ao a	em relaç ˜ao a	por causa de	em relaç ˜ao a	por causa de
de acordo com	de acordo com	em relaç ˜ao a	de acordo com	em relaç ˜ao a
por causa de	em relaç ˜ao de	de acordo com	em relaç ˜ao de	de acordo com
no final de	de acordo de	ao lado de	de acordo de	ao lado de
em torno de	com base em	em torno de	com base em	em torno de
no o in´ıcio de	em frente a	com base em	em frente a	com base em
no caso de	em direç ˜ao a	ao contr´ario de	em direç ˜ao a	ao contr´ario de
ao lado de	com relaç ˜ao a	ao longo de	com relaç ˜ao a	ao longo de
com base em	com base de	por volta de	com base de	por volta de
ao contr´ario de	em meio a	no caso de	em meio a	no caso de
ao governo de	em entrevista a	por meio de	em entrevista a	por parte de
em vez de	em contato com	por parte de	em contato com	por meio de
ao longo de	por causa de	desde o in´ıcio de	por causa de	desde o in´ıcio de
por volta de	em homenagem a	por falta de	em homenagem a	por falta de
no mercado de	no final de	de causa de	no final de	de causa de
na hora de	de combate a	em vez de	de combate a	em vez de
no centro de	em torno de	no final de	em torno de	uma vez por
na noite de	de volta a	a respeito de	de volta a	no final de
por parte de	no in´ıcio de	em acordo com	no in´ıcio de	a respeito de
em funç ˜ao de	em frente de	ao governo de	no combate a	em acordo com

Table 3: Top–ranked expressions

The alternative was to obtain a good list of CPs, compiled by NILC[1], which is used for tagging the corpus itself. The list was enriched with some "em N de" CPs, obtained from [Oliveira et al., 2003]. This manually compiled list of CPs (henceforth, the CPList) contains 169 CPs, of which 159 occurred in the corpus.

The comparison between CPList and the ranking of the statistical collocations obtained can be seen in table 4. We have found that, in the first hundred best ranked collocations of the log-likelihood test with bigrams "PRP_N PRP", there were 27 CPs from CPList, corresponding to 17.5% of the list. The same test, with bigram "PRP N_PRP", produced a better result of 22%. From this point of view, the best performing test was MI with bigrams "PRP_N PRP" which produced 86% of CPList's CPs, in the range of ten thousand best ranked collocations candidates.

The statistical tests which have been used suggest that there a lot more collocations of the pattern "PRP (Det)? N PRP" than recognisable CPs. If this is really the case then these tests have limited usefulness in building an electronic dictionary.

On the other hand, the tests can be used as facilitators. Given that above 80% of CPList's CPs were spotted within the top 10,000 collocation candidates, it is reasonable to have the 10,000 expressions analysed in the original paragraphs, by a group of human judges, in order to discover more CPs.

We feel that the statistical tests should reflect more closely the linguistic criteria. In particular, the insertion criteria could be used in a rule. For instance, if a potential CP "PRP_1 N PRP_2" occurs also as "PRP_1 Det N PRP_2", it is less likely to be a CP. If the insertion criteria is generalised as a non–modifiability criteria, then the number of variants

[1]The list was obtained from http://www.nilc.icmc.usp.br/nilc/TagSet/locucoesprepositivas.htm, in March 2004.

	LL		χ^2		MI	
Ranking	P_N P	P N_P	P_N P	P N_P	P_N P	P N_P
up to 100	27 17.5%	37 23.5%	5 3%	5 3%	27 17.5%	37 23.5%
up to 300	46 29%	65 41%	9 5.5%	26 16.5%	46 29%	65 41%
up to 500	58 36.5%	78 49%	16 10%	40 25%	58 36.5%	76 48%
up to 1.000	82 51.5%	97 61%	25 15.5%	67 42%	92 58%	92 58%
up to 10.000	124 78%	130 82%	88 55.5%	117 73%	137 86%	133 83.5%
beyond 10.000	159	159	159	159	159	159

Table 4: CPs from CPList found in the tests.

of the potential CP expressions should count as a negative factor in the statistical scores. Let us consider example 4.1

4.1 *As declarações foram feitas em entrevista* **na varanda da** *casa em que morou o presidente João Goulart.*

The expression **na varanda de** ("in the varanda of") has the desired pattern, but the fact that it has several variants in the corpus, as shown in 4.2 (insertion of the adjective **larga**) and 4.3 (plural inflexion), should be an evidence against CP–hood.

4.2 *Mas teve a compensação de ver, ao lado do seu homólogo,* **na larga varanda da** *pousada, os primeiros veículos não oficiais a atravessarem a nova ponte.*

4.3 *Os "sem convite" permaneceram atrás das cadeiras e* **nas varandas dos** *três pisos do "shopping".*

Another set of linguistic motivated test could be devised to verify whether the structure of the prepositional phrase is *$Prep_1$ NP_1(N $Prep_2$ NP_2)*, rather than *CP($Prep_1$ N_1 $Prep_2$) NP*. If *N $Prep_2$ NP_2* is found to be a collocation then the expression $Prep_1$ N_1 $Prep_2$ is less likely to be a CP. Example 4.4 illustrates this point, if we consider that **bandeira do Brasil** ("Brazilian national flag") is a MWE, which is an indication that **da bandeira de** is not a CP.

4.4 *A Prefeitura de Rio Branco distribuiu 500 kg de cal para os moradores pintarem as ruas com as cores* **da bandeira do Brasil**.

The morphology of the noun can also be used to improve the precision of the statistical methods. Let us consider, for instance, the nominalisation of the verb **extrair** ("to extract") and the impact on its complements in examples 4.5 and 4.6.

4.5 *Família e amigos do maestro comentaram sobre um possível erro de avaliação médica ao submeter Jobim a nova cirurgia, na terça-feira, para retirada do tecido ao redor de onde foi* **extraído o tumor**.

4.6 *Pessoas com câncer de pulmão avançado que são tratadas com drogas antes e depois* **da extração do tumor** *podem viver até seis vezes mais, diz um estudo dos EUA.*

While the verb complement is NP(**o tumor**), the complement of the nominalisation is a PP(**do tumor**), given rise to the potential CP **da extração do**. In summary, if we identify the morphological marks of nominalisation (i.e. derivational suffixes) in the noun then this should be negative evidence of CP–hood.

5. Concluding Remarks

Multi–word expressions have been identified statistically with success, rendering collocation tests a useful tool for building electronic lexicons. Nevertheless, the statistical methods discussed in section 3 have only limited success in identifying complex prepositions.

On the other hand, the linguistic criteria presented in section 2 is not immediately translatable into computer algorithms. They are useful for systematic human evaluation of potential CPs in sentences. Combining both tools provides a feasible method for compiling provisionary list of CPs to be used in computer applications

We suspect that there are ways in which the statistical tests could be improved, by using linguistic knowledge. Some additional filtering of the data involving the insertion criteria seems to be useful. In addition, we observed that each test produces a different ranking, as it should be expected. It would be useful to combine the tests and see if the results would improve.

References

Banerjee, S. and Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City.

Bouma, G. and Villada, B. (2002). Corpus-based acquisition of collocational prepositional phrases. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University.

CETENFolha (2004). In the WWW. http://acdc.linguateca.pt/cetenfolha.

Dias, M. C. (2002). Locução para quê? *Revista Veredas*.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mattoso Câmara Jr, J. (1984). *Dicionário de Lingüística e Gramática*. Editora Vozes.

Oliveira, C., Garrão, M., and Amaral, L. A. (2003). Recognising complex prepositions prep+n+prep as negative patterns in automatic term extraction from texts. In *Anais do I Workshop em Tecnologia da Informação e Linguagem Humana*, São Carlos, SP.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1978). *A Grammar of Contemporary English*. Longman Group Limited.

Trawinski, B. (2003). The syntax of complex prepositions in german: An hpsg approach. In Banski, P. and Przepiorkowski, A., editors, *Generative Linguistics in Poland: Morphosyntactic Investigations*, pages 155–166, Warsaw, Poland. Instytut Podstaw Informatyki Polskiej Akademii Nauk.

O Problema da Ambigüidade Lexical de Sentido na Comunicação Multilingüe

Lucia Specia, Maria das Graças Volpe Nunes

Instituto de Ciências Matemáticas e de Computação – USP
Av. do Trabalhador São-Carlense, 400, Caixa Postal 668 – 13.560-970 – São Carlos – SP
{lspecia,gracan}@icmc.usp.br

Abstract. *In this paper it is presented a discussion about the problem of word sense ambiguity in computational systems aiming at multilingual communication, specially at machine translation from English into Brazilian Portuguese. In order to do this, examples of sentences showing the implications of this linguistic phenomenon and the way it is addressed by different machine translators are analyzed. The goal is to demonstrate the importance of such module in an English-Portuguese multilingual communication system and then justify the purpose of a word sense disambiguation model between these two languages.*

Resumo. *Neste artigo é apresentada uma discussão sobre o problema da ambigüidade lexical de sentido em sistemas computacionais voltados para a comunicação multilingüe, em especial, para a tradução automática do inglês para o português do Brasil. Para tanto, são analisados exemplos de sentenças que apontam as implicações desse fenômeno lingüístico e o tratamento a ele dispensado por diferentes tradutores automáticos. Pretende-se, com isso, demonstrar a importância de um módulo dessa natureza em um sistema de comunicação multilingüe inglês-português, para então justificar a proposta de um modelo de desambiguação lexical de sentido entre essas duas línguas.*

1. Introdução

A comunicação multilingüe tem se tornado uma tarefa cada vez mais imperativa no cenário atual de grande disseminação de informações em diversas línguas, especialmente por meio da Internet. Nesse contexto, são de grande relevância os sistemas computacionais que auxiliam tal comunicação, automatizando-a, agregando a ela velocidade e praticidade. Dentre esses sistemas estão os de tradução automática, de recuperação de informações multilingües, de categorização de textos multilingües, etc.

Este trabalho está focalizado na Tradução Automática (TA), uma vez que essa aplicação representa uma etapa fundamental para a concretização de outras aplicações de comunicação multilingüe, as quais utilizam, em algum momento, um módulo de tradução. Por exemplo, um sistema de recuperação de informações multilingües normalmente emprega um módulo dessa natureza para identificar a tradução correta de um dado termo de busca nas diversas línguas envolvidas nessa busca e, com base nessa tradução, recuperar os documentos, em todas as línguas, que estão relacionados ao termo.

Muito embora seja uma das áreas mais antigas do Processamento de Línguas Naturais (PLN), a TA ainda apresenta muitos problemas. A maior parte desses problemas

está relacionada à ambigüidade entre as línguas. A ambigüidade interlingual ocorre também na tradução humana, mas é um problema especialmente grave na TA, uma vez que não se pode contar, nesse caso, com o conhecimento e a experiência da língua e do mundo do interpretador humano. A ambigüidade na tradução pode ser verificada nos diversos níveis de interpretação das línguas, incluindo o lexical, sintático, semântico, contextual e pragmático.

Este trabalho concentra-se na ambigüidade no nível lexical, que ocorre quando da multiplicidade de opções, durante a seleção de uma palavra, na língua-alvo (LA), para traduzir uma palavra da língua-fonte (LF). Ambigüidades dessa natureza caracterizam sempre uma escolha imprescindível e cujos efeitos podem ser extremamente prejudiciais à tradução, uma vez que diferentes opções podem dar origem a proposições semanticamente muito distintas. Esse problema se mostra ainda mais grave e de solução mais complexa quando são identificadas apenas variações de significado (de sentido) nas opções de tradução na LA, ou seja, quando essas opções são todas da mesma categoria gramatical. Essa variação do problema é chamada de ambigüidade lexical de sentido (ALS) e representa o foco deste trabalho, em oposição à ambigüidade categorial. A área que se ocupa do seu tratamento é denominada Desambiguação Lexical de Sentido (DLS), do inglês *Word Sense Disambiguation*.

A ALS é causada, fundamentalmente, pela existência de algumas relações semânticas interlexicais entre as línguas, principalmente a homonímia e a polissemia. Em uma das possíveis distinções entre essas duas relações, considera-se que, na **polissemia**, para uma determinada palavra da LF, existem duas ou mais palavras correspondentes na LA, com diferentes significados relacionados entre si. Por exemplo, à palavra do inglês *know* podem corresponder pelo menos duas palavras relacionadas no português, "saber" e "conhecer". Já na **homonímia**, para uma dada palavra da LF correspondem duas ou mais palavras da LA, com diferentes significados, mas não relacionados entre si. Por exemplo, a palavra do inglês *light* pode ser traduzida como "leve" ou "luz", entre outras opções. Neste trabalho, tal diferenciação não é relevante, pois pretende-se analisar ambos os fenômenos indistintamente.

Para ilustrar como a ALS é comum na tradução, considere a sentença abaixo e a quantidade de possíveis traduções (para o português) de cada uma das suas palavras de conteúdo, indicada entre parênteses, com base apenas nas traduções denotativas encontradas no dicionário eletrônico DTS DIC Prático Michaelis® 5.1. São consideradas, aqui, somente as traduções já na categoria gramatical adequada, ou seja, não há ambigüidade categorial.

"I expect (7) some (3) take (110) the veil (8) simply to hide (5) a flat (24) chest (11)".

A média de possíveis traduções para as palavras analisadas dessa sentença é 24. Se for considerado, ainda, que o significado de cada palavra pode depender do significado das demais palavras na sentença, o número de combinações possíveis é de aproximadamente 24^7.

Para realizar a desambiguação de maneira automatizada, os sistemas de TA devem incorporar um módulo de DLS ao processo de tradução. Várias abordagens têm sido propostas para a criação de módulos de DLS. Contudo, essas abordagens, na sua maioria, não são empregadas na TA, mas sim em aplicações monolingües, as quais apresentam características bastante distintas das multilingües no que se refere à manipulação da ambigüidade. Em se tratando de ambientes multilingües envolvendo o português, em especial, não se tem conhecimento de módulos de DLS desenvolvidos e efetivamente empregados.

O objetivo deste trabalho é mostrar que a ALS é um problema grave para a

comunicação multilingüe e como a falta de mecanismos de DLS afeta negativamente essa comunicação, considerando a TA inglês-português como cenário. Para tanto, são apresentados três estudos realizados com base em diversos tradutores inglês-português e sentenças de diferentes gêneros e domínios. Atenção especial é dispensada à investigação do problema da ALS na tradução de verbos, entretanto, também são apresentados resultados da análise do problema em outras classes de palavras. Este trabalho também realiza uma investigação sobre as pesquisas na área de DLS para a comunicação multilingüe, incluindo a TA, visando mostrar que não há propostas efetivas, ainda que teóricas, envolvendo a língua portuguesa.

A partir da análise do problema da ALS nos diversos sistemas de TA e da constatação de que não há abordagens para a DLS multilingüe que considerem a língua portuguesa, pretende-se, posteriormente, justificar e embasar a proposta um modelo de DLS a ser empregado em sistemas de TA do inglês para o português do Brasil.

Para mostrar as implicações do problema da ALS na tradução e o comportamento dos sistemas de TA diante desse fenômeno, na Seção 2 são apresentados os três estudos sobre a ocorrência desse problema nos diferentes sistemas. As abordagens existentes para o problema da ALS voltadas para a comunicação multilingüe são ilustradas na Seção 3. Algumas considerações e possíveis direcionamentos desse trabalho são discutidos na Seção 4.

2. O problema da ALS na TA

Os problemas causados pela ALS na tradução envolvendo o português do Brasil foram recentemente analisados em três estudos experimentais. O primeiro estudo, com base no qual este trabalho está especialmente fundamentado, consistiu da realização de um experimento com o córpus BNC (*British National Corpus*) (Burnard 2000) com o objetivo de investigar as conseqüências da ALS em traduções automáticas de textos reais, a fim de delimitar a proposta de um modelo de DLS aos casos mais problemáticos de ambigüidade[1]. Esta atividade foi desempenhada com base em três sistemas de TA inglês-português comumente utilizados, a saber, Systran, FreeTranslation e Globalink Power Translator Pro. Foram considerados para análise somente os verbos das sentenças, inicialmente, o subconjunto dos 15 verbos mais freqüentes do BNC[2]. Para a análise, 531 sentenças do BNC contendo esses 15 verbos foram aleatoriamente selecionadas e submetidas aos tradutores. As traduções foram, então, manualmente analisadas para verificar a ocorrência da ALS, seus efeitos na tradução das sentenças e o comportamento dos sistemas diante desse fenômeno.

Nesse estudo foram definidos critérios específicos para identificação de um subconjunto de verbos mais problemáticos com relação à ocorrência de ALS e à ineficiência no tratamento dispensado a ela pelos sistemas de TA. Com base nesses critérios foram selecionados sete verbos: *to go*, *to get*, *to make*, *to take*, *to come*, *to look* e *to give*. Alguns exemplos de casos de ALS encontrados no uso desses verbos e não manipulados adequadamente pelos tradutores avaliados são ilustrados na Tabela 1.

A partir desse estudo, pôde-se perceber que a porcentagem de sentenças nas quais

[1] Esse experimento é apresentado com detalhes em Specia and Nunes (2004).

[2] Essa categoria gramatical foi escolhida porque os verbos são altamente ambíguos e porque da sua desambiguação pode depender a desambiguação de outras palavras da sentença, principalmente dos seus argumentos. Posteriormente, pretende-se estender esse trabalho a outras categorias lexicais.

ocorre ambigüidade nos sete verbos selecionados é bastante grande. De acordo com os critérios definidos, foram consideradas sentenças problemáticas somente aquelas cuja acepção do verbo em foco não era corretamente identificada por pelo menos dois sistemas. Com base nesse critério, das 238 sentenças com os sete verbos, 149 foram consideradas problemáticas (62,6% do total). Se fossem consideradas problemáticas as sentenças nas quais a acepção correta do verbo não havia sido identificada por pelo menos um sistema, esse número aumentaria para 177 sentenças (74,4% do total). Esse número alto mostra que os sistemas estudados não dispõem de mecanismos de DLS. Normalmente, eles escolhem uma das possíveis acepções de um verbo, provavelmente a mais comum, e essa acepção é utilizada na tradução da maioria das suas ocorrências, excetuando-se alguns casos do uso do verbo em *phrasal verbs* ou em expressões comuns. O tratamento dispensado a *phrasal verbs* é também bastante simplificado: muitas vezes, um verbo seguido de uma preposição (dois elementos que poderiam compor um *phrasal verb*) é diretamente traduzido como o *phrasal verb* correspondente, mesmo que não seja usado com tal função na sentença, como ocorre com a última sentença da Tabela 1.

Tabela 1. Exemplos de sentenças do BNC com verbos problemáticos

Sentença	Acepção correta	TA		
		Systran	Free-Translation	Power Translator
The war may well just **go** on and on.	continuar	ir	vai	ir
Stand in a French village when the Tour de France **goes** by and you are participating in an event which is unambiguously French.	passa (passar)	vai	vai	passa
It's best to be alone when the noises **get** this loud.	ficam (ficar)	recebem	começam	adquirem
A lot of international help will be needed to **get** things moving.	fazer	receber	começar	adquirir
They **take** more foreign holidays.	têm (ter)	tomam	fazem exame	levam
"**Take** that money out of your mouth!" said her mother.	tire (tirar)	toma … fora	faça exame … fora	objeto pega-do … fora
Now eat your supper, both o' ye, afore it **takes** cold.	fique (ficar)	toma	faz exame	leva
"This city has suddenly **come** alive," said her husband, an off-duty border guard.	renasceu (renascer)	veio vivo	vivo ... vindo	veio viva
"Yes, I'm **coming**, but I've one or two things to attend to first," she explained.	indo (ir)	venho	vindo	vindo
Mr Gonzalez has also **come** in for criticism from within his own party.	recebeu (receber)	entrou	entrou	entrou

O segundo estudo (Fossey et al. 2004) foi desenvolvido como parte da avaliação do sistema de TA inglês-português EPT-Web[3], ainda em construção. Nele, foi analisada, entre outros problemas, a ocorrência da ambigüidade lexical na TA inglês-português. Para tanto, foi considerado um córpus de textos do jornal *New York Times* (NYT) *on-line* e quatro ferramentas de tradução disponíveis na web: Linguatec E-Translation Server, Intertran, Systran e FreeTranslation. Esse estudo considerou indistintamente ambos os tipos de ambigüidade lexical, isto é, categorial e de sentido (homonímia e polissemia). Foram avaliadas

[3] http://www.nilc.icmc.usp.br/nilc/projects/ept-web.htm

as traduções de todas as palavras de conteúdo de 515 sentenças, nos quatro sistemas. Uma sentença foi considerada problemática em um sistema se apresentasse pelo menos uma palavra ambígua inadequadamente traduzida por esse sistema. Os números e percentuais de sentenças problemáticas, em cada tradutor, são apresentados na Tabela 2.

Tabela 2. Sentenças do NYT com ambigüidade lexical

Sistema	Nº de sentenças cuja acepção não foi corretamente identificada	% de sentenças cuja acepção não foi corretamente identificada
E-Translation	279	54,1
Intertran	361	70,1
Systran	272	52,8
FreeTranslation	271	52,6

No estudo também foram apresentados os percentuais de palavras ambíguas cuja acepção não foi corretamente identificada pelos tradutores, agrupadas de acordo com a sua categoria gramatical, com relação ao total de palavras ambíguas. Nos quatro sistemas avaliados, a maioria das palavras ambíguas se distribuía entre substantivos e verbos, conforme ilustrado na Tabela 3. Exemplos de sentenças com problemas de tradução causados pela ALS apenas dos verbos em alguns sistemas são ilustrados na Tabela 4.

Tabela 3. Verbos e substantivos ambíguos do NYT

Sistema	% de substantivos cuja acepção não foi corretamente identificada	% de verbos cuja acepção não foi corretamente identificada
E-Translation	36,7	29,8
Intertran	38,7	32,3
Systran	39,6	24,1
FreeTranslation	40,0	31,4

Tabela 4. Exemplos de sentenças do NYT com ALS nos verbos

Sentença	Acepção correta	TA	Sistema de TA
With an Organic Sensor, a Food Wrapper **Sniffs Out** Trouble.	descobre (descobrir)	funga fora	E-Translation
Bush **Sending** Powell to Middle East.	enviando (enviar)	emite	Systran
Click Here to Receive 50% Off Home Delivery of The New York Times Newspaper.	clique (clicar)	estale	Systran
Check them **out**, or post any wine-related topics.	dê baixa (dar baixa)	verifique-os para fora	FreeTranslation

Pela Tabela 2, pode-se observar que os quatro sistemas apresentaram um porcentual maior que 50% de sentenças com problemas específicos de ambigüidade lexical, em uma ou mais palavras. Em uma análise realizada, ainda no estudo citado, sobre a gramaticalidade das sentenças, foi verificado que a maior parte de sentenças problemáticas correspondem a sentenças agramaticais ou gramaticais com tradução incorreta. Com isso, em alguns casos, ainda que os problemas de ambigüidade lexical fossem resolvidos, as sentenças permaneceriam incorretas, já que eram agramaticais, mas, em grande parte dos casos, uma vez resolvido o problema da ambigüidade, as sentenças poderiam, em sua maioria, tornar-se

semanticamente corretas. Assim, segundo os autores, fica evidente que a ambigüidade lexical compromete profundamente a qualidade das traduções produzidas automaticamente e que a solução das questões envolvendo esse problema se mostra um dos caminhos necessários para a obtenção de resultados mais satisfatórios nas produções das ferramentas de TA.

O terceiro estudo foi realizado por Oliveira et al. (2000). Os autores analisaram, comparativamente, vários sistemas de TA entre inglês e português, comerciais ou disponíveis na web, avaliando ambas as direções da tradução. Para testar o desempenho dos sistemas na direção inglês-português, 10 passagens de textos do jornal *New York Times* (com uma ou mais sentenças, totalizando 530 itens lexicais) foram submetidos a cinco sistemas: Globalink Power Translator Pro, Alta Vista, Intertran, Tradunet e Linguatec E-Translation Server. As traduções foram analisadas para identificar problemas em três níveis de interpretação: lexical, sintático e semântico-pragmático.

No nível lexical, o desempenho dos sistemas foi testado em quatro situações: dicionarização, ambigüidade, conotação e expressões idiomáticas. No caso das ambigüidades lexicais, foram consideradas, indistintamente, as ambigüidade categorial e de sentido (polissemia ou homonímia). Alguns exemplos de problemas causados pela ALS na tradução, selecionados entre os relatados, são ilustrados na Tabela 5.

Tabela 5. Exemplos de sentenças com ALS (Oliveira et al. 2000)

Sentença	Acepção correta	TA	Sistema de TA
(...) Hungary has ceded more sovereignty than many other nations – including the United States – would **ever** consider (...)	sempre	jamais	Translator Pro
To paraphrase a celebrated epitaph, prosperity left **scarcely** any of our industries untouched, and touched nothing it did not enrich.	mal	quase nenhuma	E-Translation

Segundo os autores, a presença da ambigüidade lexical na TA entre o inglês e o português é bastante freqüente, justificando a necessidade de estratégias de desambiguação nas ferramentas de tradução. Eles afirmam que a qualidade das escolhas lexicais afeta o processo de tradução em vários graus, principalmente se a escolha incorreta ocorrer em itens lexicais em posições de núcleo, como verbos em um predicado verbal ou substantivos em um sujeito. Nesses casos, a ambigüidade lexical pode prejudicar a coerência local e global da sentença, freqüentemente tornando-a incompreensível.

Nesse estudo, os autores também verificaram que as ferramentas de tradução não empregam mecanismos para procurar resolver o problema da ambigüidade lexical. Em vez disso, apostam em decisões baseadas em critérios muito simples, como a freqüência da ocorrência de cada acepção em traduções reais. A maioria dos erros encontrados, segundo os autores, diz respeito a expressões com grupos de palavras que podem assumir significados diferentes da composição do significado que elas possuem individualmente, como ocorre, por exemplo, em *phrasal verbs*. A conclusão geral dos autores é que a qualidade das traduções poderia ser consideravelmente aprimorada se fosse assumida uma perspectiva diferente com relação às idiossincrasias de cada língua, ou seja, se fossem empregados esforços de caráter mais efetivo para o tratamento dessas idiossincrasias.

De modo geral, apesar de terem objetivos distintos, os três estudos citados

apresentam resultados que corroboram a hipótese de que a ambigüidade lexical influencia negativamente nos resultados da comunicação multilingüe, em especial, na TA inglês-português, e que mostram que esse problema não recebe, ainda, tratamento adequado nas ferramentas disponíveis. Com isso, comprovam a necessidade de mecanismos de DLS para essa comunicação.

3. Abordagens para a DLS na comunicação multilingüe

Várias abordagens de DLS têm sido propostas para diversas aplicações, principalmente para aquelas monolingües. Essas abordagens podem seguir diferentes métodos de PLN: **métodos lingüísticos**, baseados em conhecimento lingüístico e/ou extralingüístico explicitamente especificado, manualmente ou semi-automaticamente, por meio de recursos como dicionários eletrônicos; **métodos empíricos**, baseados em córpus de exemplos e em algoritmos de aprendizado de máquina para adquirir conhecimento automaticamente a partir dos exemplos; ou **métodos híbridos**, que combinam características dos métodos lingüísticos e empíricos.

Considerando-se a aplicação específica da DLS em tarefas multilingües, são poucos os trabalhos desenvolvidos de que se tem conhecimento. No caso de abordagens que seguem métodos lingüísticos, pode-se citar os trabalhos de Egedi et al. (1994), Dorr and Katsova (1998), Pedersen (1997) e Montoyo et al. (2002). Egedi et al. (1994) apresentam um sistema de TA do coreano para o inglês que possui um módulo de DLS para tratar da polissemia de alguns verbos, com base na unificação de restrições de seleção semânticas definidas na estrutura argumental desses verbos com os traços semânticos definidos para os substantivos que podem ser utilizados como seus argumentos. As restrições de seleção e traços semânticos são especificados na LA. Assim, a desambiguação de um verbo depende da tradução correta dos seus argumentos. Certamente, tal abordagem apresentará problemas se os argumentos do verbo também forem ambíguos.

Dorr and Katsova (1998) definem um mecanismo de seleção lexical para verbos e substantivos deverbais que se baseia na estrutura argumental desses elementos, representada por meio de Estruturas Conceituais Lexicais, e nos sentidos da WordNet. A hipótese é de que a tradução de um elemento da LF pode ser desambiguada se forem escolhidos, na LA, elementos que apresentem a mesma LCS e que estejam no mesmo *synset* da WordNet, ou seja, que sejam sinônimos do elemento na LF. Em experimentos considerando a desambiguação do inglês para o espanhol, as autoras obtiveram resultados promissores, constatando que os elementos são facilmente desambiguados, pois são raros os elementos com a mesma LCS que são sinônimos. Contudo, o mecanismo proposto exige uma base de dados com todos os itens lexicais representados por estruturas LCS e previamente mapeados (manualmente) em um *synset* da WordNet.

Montoyo et al. (2002) discutem a necessidade de um módulo de DLS em aplicações multilingües voltadas para recuperação de informações, e apresentam uma interface para a desambiguação de substantivos e verbos que poderia ser acoplada a esses sistemas. Nessa interface, consideram o espanhol e o inglês como LF, e a taxonomia da EuroWordNet para realizar o mapeamento entre as palavras dessas duas línguas e também o mapeamento para o catalão e o basco. A desambiguação realizada consiste, basicamente, em identificar qual é o código da EuroWordNet correspondente à palavra a ser desambiguada, ainda na LF e, em seguida, encontrar a palavra na LA com o mesmo código da EuroWordNet. Assim, embora seja voltada para aplicações multilingües, a desambiguação é feita de maneira monolingüe.

Essa abordagem só se mostra viável para línguas previstas no projeto EuroWordNet, para as quais já existem códigos correspondentes aos itens lexicais.

Pedersen (1997) descreve um mecanismo para a desambiguação de um subconjunto de verbos de movimento polissêmicos na TA do dinamarquês para o inglês. A autora considera apenas a polissemia sistemática desse subconjunto. Para tanto, utiliza a abordagem de esquemas para descrever os verbos a partir de uma grande quantidade de informações lingüísticas da LF, em diversos níveis, para auxiliar na desambiguação. Apesar da aplicação voltada para a TA, o foco da autora é na especificação desses esquemas com informações suficientes para permitir capturar os padrões sistemáticos entre os diferentes sentidos de um verbo, de modo a evitar descrições ambíguas. Assim, a desambiguação ocorre, em grande parte, na LF. Além disso, a estrutura definida para os esquemas é específica para verbos de movimento do dinamarquês, o que dificulta extensões dessa abordagem.

No caso de abordagens que seguem métodos empíricos, podem ser considerados os trabalhos de Brown et al. (1991) e Lee (2002). Brown et al. (1991) descreve uma abordagem estatística para a seleção lexical na TA entre o francês e o inglês. Essa abordagem é bastante simples, pois deriva do teorema de Bayes, que se baseia principalmente na freqüência de cada possível tradução em um córpus. Além disso, ela considera uma desambiguação binária, ou seja, a escolha entre apenas dois possíveis sentidos de uma palavra ambígua. O módulo de desambiguação desenvolvido foi avaliado em um sistema de TA, também estatístico, considerando as 500 palavras mais comuns do inglês e as 200 mais comuns do francês. Com esse módulo, a taxa de erro nas traduções resultantes do sistema, segundo os autores, diminuiu 13%. Contudo, é preciso levar em conta as características limitadas do contexto de desenvolvimento desse modelo.

Lee (2002) apresenta uma abordagem baseada em córpus para a seleção lexical na TA inglês-coreano. Essa abordagem trata a seleção lexical como um problema de classificação e emprega um algoritmo para a escolha pela classe (tradução) mais adequada. Tal algoritmo também é estatístico, mas contempla outras características, alem da freqüência das traduções. O autor tem por objetivo obter um modelo portável, independente de língua. Para tanto, utiliza como características para a classificação somente as outras palavras da sentença em que a palavra ambígua ocorre, agrupadas de duas a duas. Com isso, a precisão da classificação gerada, quando o modelo é avaliados com novos casos, é pouco melhor que a da *baseline* considerada (neste caso, a tradução mais freqüente). O autor não cita a abrangência do modelo, mas como esse modelo se baseia nas palavras da sentença, tal abrangência provavelmente é bastante limitada.

Considerando-se as abordagens de DLS para a TA que seguem métodos híbridos, pode-se citar o trabalho de Zinovjeva (2000). Esse trabalho tem por objetivo aprender automaticamente regras de transformação para traduzir corretamente palavras ambíguas do inglês para o sueco. Para tanto, utiliza conhecimento pré-codificado em recursos lingüísticos como dicionários, por meio de procedimentos, tais como um etiquetador morfossintático, que atuam como filtros, eliminando alguns dos possíveis sentidos de cada palavra ambígua. O resultado é um conjunto reduzido de sentidos para cada palavra ambígua, em um determinado contexto. Para a geração do modelo de DLS, são fornecidos exemplos de desambiguação a um algoritmo de aprendizado de máquina, baseado no método de aprendizado por transformações. Esses exemplos são formados por sentenças com palavras ambíguas, seu contexto e a suas correspondentes traduções, identificadas manualmente.

Gerado o modelo, para desambiguar um novo caso, os resultados dos procedimentos preliminares são fornecidos ao modelo, que escolhe pela tradução mais adequada. Apesar de ser considerada uma abordagem híbrida, a utilização do conhecimento pré-codificado e a aquisição de novos conhecimentos (ou seja, das regras de transformação) ocorrem em etapas isoladas no processo de DLS, assim, pouco se aproveita de todo o potencial das metodologias híbridas.

Os trabalhos citados, apesar de não representarem um lista exaustiva de todos os encontrados, exemplificam a maioria deles, mantendo a proporção da distribuição entre os tipos de abordagens. Como se pode perceber, a maioria desses trabalhos é baseada em métodos lingüísticos. Apesar de poderem apresentar resultados bastante precisos, as dificuldades para a criação das fontes de conhecimento acabam restringindo muito a abrangência desses trabalhos. Abordagens baseadas em métodos empíricos permitem modelos mais abrangentes, mas são ainda pouco pesquisadas no contexto da TA, em função da dificuldade na criação de córpus de exemplos representativos e consistentes. Abordagens seguindo métodos híbridos, por sua vez, permitem combinar as vantagens de ambos os métodos, mas são raros os trabalhos desenvolvidos sob esse método.

Pode-se observar, também, que nenhum dos trabalhos citados inclui desambiguações para o português. No único trabalho envolvendo o português encontrado, Leffa (1998) cita a importância do uso do contexto local da palavra ambígua, isto é, das palavras vizinhas a ela na sentença, para a desambiguação na TA. Em um experimento para desambiguar 20 palavras ambíguas do inglês para o português, o autor relata um desempenho de 94%. No entanto, como esse modelo se baseia nas palavras da sentença, sua abrangência deve ser bastante limitada. Além disso, o trabalho foi interrompido.

4. Considerações finais

Neste trabalho foram relatados os resultados de alguns estudos experimentais com diversos sistemas de TA inglês-português, focalizando o problema da ALS. Por meio desses estudos foi possível constatar que a ALS é um problema bastante comum, proeminente e prejudicial para a tradução. Além disso, foi possível verificar que todos os sistemas testados, de uso expressivo atualmente, não oferecem tratamento adequado para esse problema. Em se tratando da TA para o português, particularmente, de fato, não se tem conhecimento de sistemas que empreguem mecanismos de DLS. Pode-se concluir, com isso, que a falta desses mecanismos é certamente um dos principais motivos para os resultados bastante insatisfatórios dos sistemas existentes.

Considerando-se as pesquisas teóricas, foi possível verificar que há poucos trabalhos especificamente voltados para a TA e que as metodologias normalmente exploradas não permitem criar abordagens com resultados precisos e, ao mesmo tempo, abrangentes. Além disso, não se tem conhecimento de trabalhos significativos envolvendo o português.

A partir dos resultados dessa análise sobre o problema da ALS na TA inglês-português e da investigação das abordagens existentes para esse problema (para outras aplicações ou envolvendo outras línguas), pretende-se desenvolver um modelo computacional de DLS voltado especificamente para a TA do inglês para o português do Brasil. Esse modelo será construído seguindo um método realmente híbrido de PLN, ou seja, será baseado em conhecimento lingüístico e em córpus, com a utilização de conhecimento substancial durante o processo de aprendizado automático. Essa configuração permitirá

que o modelo seja abrangente, aplicável a sistemas de TA em larga-escala, auxiliando no processo de escolha lexical de uma grande quantidade de palavras, e que apresente resultados potencialmente melhores que os dos trabalhos já existentes para outras línguas. Adicionalmente, esse modelo deverá ser o mais independente possível, de maneira que possa ser acoplado a diferentes sistemas de TA.

Desenvolvida uma abordagem eficiente para o problema da ALS, esta poderá representar melhorias significativas na qualidade dos sistemas de TA inglês-português atuais. Entre os sistemas que devem se beneficiar com esse módulo está o EPT-Web, ainda em desenvolvimento, para a TA de textos jornalísticos do inglês para o português do Brasil. Apesar de o foco inicial ser a TA, esse módulo poderá também ser empregado, posteriormente, em diferentes aplicações envolvendo a comunicação multilingüe.

Referências Bibliográficas

Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R. (1991) "Word sense disambiguation using statistical methods", In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, p. 264-270.

Burnard, L. (2000) "Reference Guide for the British National Corpus (World Edition)", Oxford University Press.

Dorr, B. J. and Katsova, M. (1998) "Lexical Selection for Cross-Language Applications: Combining LCS with WordNet", In: Proceedings of AMTA'1998, Langhorne, p. 438-447.

Fossey, M.F., Pedrolongo, T., Martins, R. T. and Nunes, M. G. V. (2004) "Análise comparativa de tradutores automáticos inglês-português", Série de Relatórios do NILC, NILC-TR-04-04, São Carlos, Março, 18p.

Leffa, V. J. (1998) "Textual constraints in L2 lexical disambiguation", System, Great Britain, 26(2), p. 183-194.

Lee, H. (2002) "Classification Approach to Word Selection in Machine Translation", In: Proceedings of AMTA'2002, Springer-Verlag, Berlin, p. 114-123.

Montoyo, A., Romero, R., Vazquez, S., Calle, M. and Soler, S. (2002) "The Role of WSD for Multilingual Natural Language Applications", In: Proceedings of TSD'2002, Czech Republic, p. 41-48.

Oliveira Jr., O.N., Marchi, A.R., Martins, M.S. and Martins, R.T. (2000) "A Critical Analysis of the Performance of English-Portuguese-English MT Systems", In: Anais do V PROPOR, Atibaia, p. 85-92.

Pedersen, B. S. (1997) "Lexical ambiguity in machine translation: expressing regularities in the polysemy of Danish Motion Verbs". PhD Thesis, Center for Sprogteknologi, Copenhagen, Denmark.

Specia, L. and Nunes, M.G.V. (2004) "A ambigüidade lexical de sentido na tradução do inglês para o português – um recorte de verbos problemáticos", Série de Relatórios do NILC, NILC-TR-04-01, São Carlos, Março, 30p.

Zinovjeva, N. (2000) "Learning Sense Disambiguation Rules for Machine Translation". Master's Thesis in Language Engineering. Department of Linguistics, Uppsala University.

Identificação do perfil dos usuários da Biblioteca Central da FURB através de *data mining* para a personalização da recuperação e disseminação de informações

Alberto Pereira de Jesus[1], Evanilde Maria Moser[1], Paulo José Ogliari[2]

[1]Biblioteca Central – Universidade Regional de Blumenau (FURB)
Caixa Postal 15.05 – 89.012-900 – Blumenau – SC – Brasil

[2]Departamento Informática e Estatística – Universidade Federal de Santa Catarina
{albertop,emmoser}@furb.br, ogliari@inf.ufsc.br

Abstract. *This paper describes all data mining deployment stages applied to library user profile identification for information recovery and dissemination WEB systems.*

Resumo. Este artigo descreve todas as etapas de implantação de *data mining* aplicado na identificação do perfil dos usuários de uma biblioteca para a personalização de sistemas WEB de recuperação e disseminação de informações.

1. Introdução

Com o crescimento do volume de publicações e também das necessidades de informações dos clientes, sejam elas em papel ou em formato eletrônico, é importante, que as bibliotecas possuam sistemas de informações capazes de armazenar e indexar informações bibliográficas de forma a facilitar a recuperação e disseminação aos usuários (CARDOSO, 2000).

Neste sentido, dois sistemas têm sido desenvolvidos, sendo eles o sistema de recuperação de informações (SRI) e o sistema de disseminação seletiva de informações (DSI). Enquanto o SRI trata de localizar as informações solicitadas pelo usuário, o DSI tenta prever as necessidades desses usuários, fazendo recomendações e sugestões conforme seu interesse.

Assim, conhecer os usuários é importante e já era uma necessidade no passado, onde o bibliotecário sabia e conseguia lembrar as preferências de cada um de seus usuários para fazer recomendações e ajudá-los na localização de obras. Hoje, devido à grande quantidade de usuários e publicações, precisa-se de ferramentas automatizadas que auxiliem nesse processo.

Sabendo-se que a missão das Bibliotecas, segundo Funaro et al. (2001, p. 1) "é oferecer a seus usuários informações relevantes para a realização de suas pesquisas, facilitando o acesso e localização do material necessário", os sistemas tradicionais de SRI e DSI das Bibliotecas necessitam evoluir e ser inteligentes, a fim de agregar valor ao serviço de referência. Dessa forma é necessário que se conheça o perfil do usuário, delineando suas preferências e seus interesses.

As técnicas de *data mining* permitem a identificação desse perfil, possibilitando a personalização dos processos do SRI e DSI, tornando-os objetivos e seletivos. Esta confluência de acertos caracteriza a relevância da informação. Não adianta o usuário

receber uma comunicação personalizada se ela não for relevante para seus interesses e necessidades.

"O objetivo da personalização de conteúdo é garantir que a pessoa certa receba a informação certa no momento certo" (ARANHA, 2000, p. 10).

Estes sistemas, principalmente o DSI, apesar das facilidades que oferece, apresenta alguns problemas como: o não preenchimento por alguns usuários e as rápidas mudanças que ocorrem em seus interesses. Toma-se como exemplo um professor que lecionava uma disciplina de *data warehouse* e atualmente leciona a disciplina de *data mining*. Como ele preencheu seus dados com antigo perfil, continuará recebendo informações sobre seus interesses preenchidos anteriormente.

Seria prudente que o sistema reconhecesse essas alterações no ambiente e fosse capaz de se adequar às novas características. Isso é possível por meio da aplicação de técnicas de *data mining* sobre os dados contidos nos registros de transações como: empréstimos, reservas e consultas que são armazenados no banco de dados da Biblioteca e servirão para fazer um estudo do perfil do usuário. Estes registros são armazenados diariamente pelas transações de empréstimos, no entanto não são utilizados para tomada de decisão.

Mais especificamente, a aplicação de *data mining* nestes registros permitirá:

a) melhorar o SRI através da personalização das consultas, ao fazer uma busca o retorno da consulta é filtrado segundo o perfil do usuário;
b) facilitar o DSI, recomendando obras de interesse ao usuário.

2. Objetivos

O objetivo geral deste trabalho é desenvolver um sistema de recuperação e disseminação de informações, personalizado segundo o perfil de cada usuário da Biblioteca Central (BC) da Universidade Regional de Blumenau (FURB), por meio da aplicação de técnicas de *data mining*. Como objetivos específicos têm-se:

a) desenvolver um *data warehouse* para dar suporte a aplicação das técnicas de *data mining*, possibilitando também obter informações para tomada de decisões;
b) aplicar técnicas de *data mining* sobre o histórico de empréstimos e reservas dos usuários para identificar o perfil dos mesmos na BC da FURB;
c) desenvolver um sistema WEB de SRI e DSI personalizado dinamicamente para a BC da FURB.

3. Justificativa

O trabalho se justificativa, pois conhecendo as características e preferências do usuário, pode-se assim definir seu perfil que é de elevada importância para o SRI, DSI e para tomada de decisões gerenciais. Possibilitando uma maior satisfação dos usuários, uma melhor utilização e organização da biblioteca, redução de custos com a aquisição de materiais, bem como, facilidade no atendimento dos usuários.

4. Fundamentação Teórica

A quantidade de informações produzidas versus a capacidade de armazenamento dos recursos computacionais a um baixo custo, tem impulsionado o desenvolvimento de novas tecnologias capazes de tratar estes dados, transformá-los em informações úteis e extrair conhecimentos.

Entretanto, o principal objetivo da utilização do computador ainda tem sido o de resolver problemas operacionais das organizações, que coletam e geram grandes volumes de dados que são usados ou obtidos em suas operações diárias e armazenados nos bancos de dados. Porém, os mesmos não são utilizados para tomadas de decisões, ficando retidos em seus bancos de dados, sendo utilizados somente como fonte histórica. Estas organizações têm dificuldades na identificação de formas de exploração desses dados, e principalmente na transformação desses repositórios em conhecimento (BARTOLOMEU, 2002).

Pesquisadores de diferentes áreas estudam e desenvolvem trabalhos para obter informações e extrair conhecimentos a partir de grandes bases de dados, como tópico de pesquisa, com ênfase na técnica conhecida como *data mining*.

Data mining é parte do processo de *Knowledge Discovery in Databases* (KDD), ou descoberta de conhecimentos em bancos de dados (DCBD), o qual é responsável pela extração de informações sem conhecimento prévio, de um grande banco de dados, e seu uso para a tomada de decisões (DINIZ; LOUZADA NETO, 2000).

KDD é um processo contínuo e cíclico que permite que os resultados sejam alcançados e melhorados ao longo do tempo. Na figura 1 são apresentados os passos que devem ser executados no processo de KDD. Segundo Diniz; Louzada Neto (2000) embora os passos devam ser seguidos na mesma ordem em que são apresentados, o processo é extremamente interativo e iterativo (com várias decisões sendo feitas pelo próprio usuário e loops podendo ocorrer entre quaisquer dois ou mais passos).

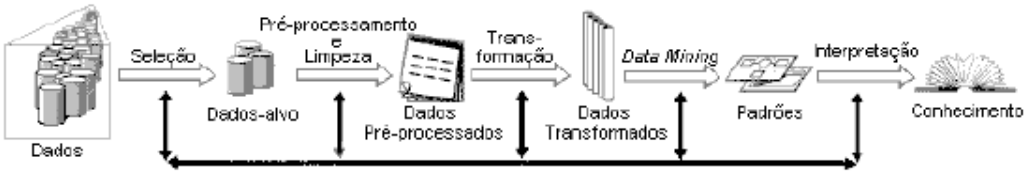


Figura 1 – Passos do processo de KDD (Fonte: FIGUEIRA, 1998, p. 8.)

Para a implantação desta tecnologia é necessário que se conheça a fundo o processo, para que a mesma venha atender às expectativas do usuário. O processo de KDD começa obviamente com o entendimento do domínio da aplicação e dos objetivos finais a serem atingidos.

Segundo Harrison (1998, p. 155) "*data mining* é a exploração e análise, por meios automáticos ou semi-automáticos, das grandes quantidades de dados para descobrir modelos e regras significativas".

Deve-se destacar que cada técnica de *data mining* ou cada implementação específica de algoritmos que são utilizados para conduzir as operações *data mining* adapta-se melhor a alguns problemas que a outros, o que impossibilita a existência de um método de *data mining* universalmente melhor. Para cada particular problema tem-se um particular algoritmo. Portanto, o sucesso de uma tarefa de *data mining* está diretamente ligado à experiência e intuição do analista (Diniz; Louzada Neto 2000).

Assim, é importante que se conheçam, as tarefas desempenhadas (Classificação, Estimatição, Previsão, Agrupamento por afinidade, Segmentação e Descrição) e suas técnicas (Análise de seleção estatística, Raciocínio baseado em casos, Algoritmos genéticos, Detecção de agrupamentos, Análise de vínculos, Árvores de decisão e indução de regras, Redes neurais) a fim de dar suporte a sua escolha.

5. Metodologia

Para o processo de extração de conhecimento nos dados da biblioteca sobre o perfil dos usuários, será utilizada a metodologia referenciada por Berry; Linoff (1997). A Figura 2 apresenta o modelo proposto. A mesma será aplicada na BC da FURB.

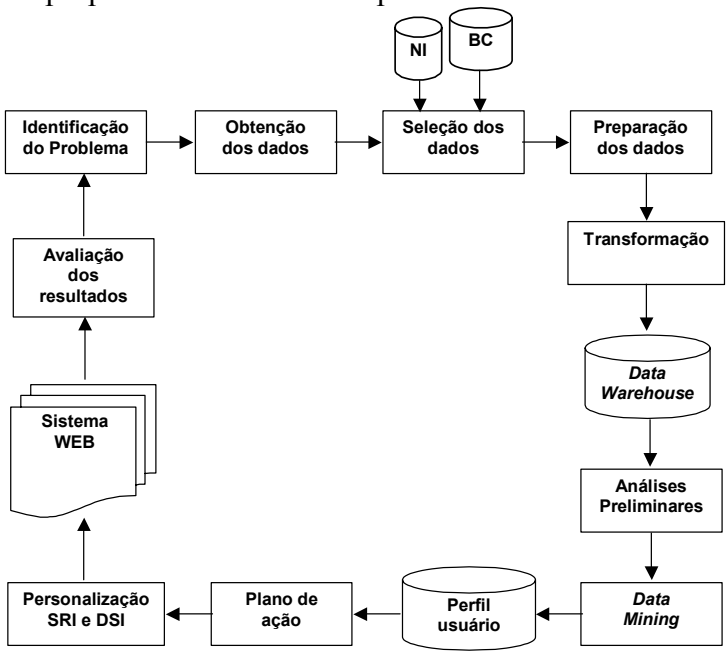


Figura 2. Modelo proposto para aplicação de *data mining* em bibliotecas

5.1. Identificação do problema

A maioria das bibliotecas não possui sistemas de recuperação e disseminação de informações capazes de ajudar no processo de localização das obras de interesse dos usuários. O mesmo é feito pelo serviço de referência com o auxílio de bibliotecários (a) ou especialistas na área.

A BC da FURB atualmente não apresenta nenhum sistema informatizado de DSI aos usuários. Este ainda é feito de forma manual pelo serviço de referência. O SRI não identifica o usuário para tratá-lo de forma seletiva e personalizada. Quando é feita uma consulta, a pesquisa retorna uma grande quantidade de informações (dados) a maioria sem relevância e nenhuma ordenação, o que caracteriza uma alta revocação, mas baixa precisão (CARDOSO, 2000, p. 2).

Definido o problema, elegem-se as variáveis que serão utilizadas na investigação para a resolução do mesmo. As variáveis que tem relacionamento direto para identificação do perfil dos usuários são: usuários; obras da Biblioteca; CDD (classificação decimal dewey do assunto principal da obra); transações (empréstimos, reservas).

5.2. Obtenção dos dados

Mediante a identificação das variáveis que serão utilizadas no processo de extração de conhecimento sobre o perfil do usuário, parte-se para o reconhecimento e a obtenção das mesmas nas fontes de dados. A principal fonte dos dados são os sistemas legados da BC mantidos pela seção de automação (cadastro e a circulação obras). Outra fonte é o sistema de identificação única de pessoas com vínculo na instituição, mantido pelo NI (usuários). A ligação entre o usuário e suas transações é feita através do identificador único do usuário.

5.3. Seleção dos dados

Através de uma engenharia reversa das tabelas de interesse armazenadas nos bancos de dados, torna-se possível reconhecer as variáveis de interesse para assim fazer a seleção.

Foram selecionados na amostra professores e alunos de pós-graduação da FURB. Com a integração das bases, excluem-se alguns dados como CPF, Endereço, etc, por serem usadas com finalidades operacionais que não se aplicam a esta pesquisa.

5.4. Pré-processamento dos dados

Após a seleção dos dados, faz-se a verificação da existência de inconsistências e/ou erros nas variáveis: A data de aquisição continha dados fora do formato padrão e o código CDD em alguns casos estava fora do padrão de catalogação.

5.5. Extração, transformação e carga dos dados

Os dados são estruturados para facilitar e agilizar o processo de mineração. A partir daí, foi gerado um *data mart*, que é parte de um *data warehouse*. Após identificar as variáveis de interesse, chega-se a um modelo que trata da circulação das obras.

A tabela fato é a de circulação de empréstimos, onde cada registro corresponde a uma transação que pode ser dos tipos: empréstimo e reserva. As dimensões encontradas são: a obra e o usuário que a emprestou. A partir do modelo de *data mart* foram criadas as tabelas e rotinas para carga dos dados.

Na área de biblioteconomia já foram institucionalizados alguns códigos para determinados domínios de uma variável, como é o caso da classificação dos livros. Existe uma codificação internacional, conhecida como Classificação Decimal Dewey (CDD), que é usada por diferentes órgãos da área de biblioteconomia, a fim de organizar o acervo e facilitar a localização das obras. Assim foram criados cinco níveis da CDD. A partir da qual foram gerados os assuntos significativos (AS) através da totalização das transações segundo a CDD, reduzindo as mesmas até um nível mínimo de significância.

5.6. Análises preliminares

Em qualquer investigação é fundamental para o pesquisador ter uma visão global dos dados que estão sendo pesquisados, a seguir apresenta-se uma análise descritiva dos dados da amostra, envolvidos neste estudo.

A amostra é composta por 17421 títulos que totalizam 51011 volumes, 3906 usuários, 821 da categoria professores e 3085 da categoria de pós-graduação.

Estes usuários realizaram 68543 transações, 66769 de empréstimo e 1775 de reservas. Tivemos uma média de 17,54 transações por usuários.

5.7. Mineração dos dados

Caracteriza-se pela transformação dos dados em conhecimento. Para encontrar o perfil do usuário utilizam-se as seguintes etapas:

5.7.1 Análise de conglomerados de assuntos significativos

A metodologia de análise de conglomerados (*cluster analysis*) é uma descoberta indireta de conhecimento a partir de algoritmos para encontrar registros de dados que são semelhantes entre si. Estes conjuntos de registros similares são conhecidos como clusters.

Segundo Velasquez et al. (2001, p. 2) “Todos os algoritmos de análise de conglomerados são baseados em uma medida de similaridade ou, ao contrário de

distância, que procuram expressar o grau de semelhança entre os objetos". Uma medida de distância muito utilizada quando os atributos são de natureza quantitativa é a distância euclidiana.

Formam-se agrupamentos das obras em grandes áreas de conhecimento, ou seja, grupos de livros os quais são utilizados por usuários para estudo de determinando assunto ou área. Assim foram analisados alguns métodos estatísticos de agrupamento hierárquico, como o do vizinho mais próximo, do vizinho mais distante, e de Ward. Optou-se pelo método Ward com distância euclidiana, pois o mesmo apresentou melhores resultados e por ser indicado por Aranha (2000) em seu trabalho.

Afirma Velasquez et al. (2001, p. 2) que "Nos métodos hierárquicos o número de classes não é fixado a priori, mas resulta da visualização do dendrograma, um gráfico que mostra a seqüência das fusões ou divisões ao longo do processo iterativo".

Foi aplicado esta técnica as dados de transações dos usuários segundo suas transações por AS, contidos no *data mart* resultando no dendograma apresentado na Figura 3.

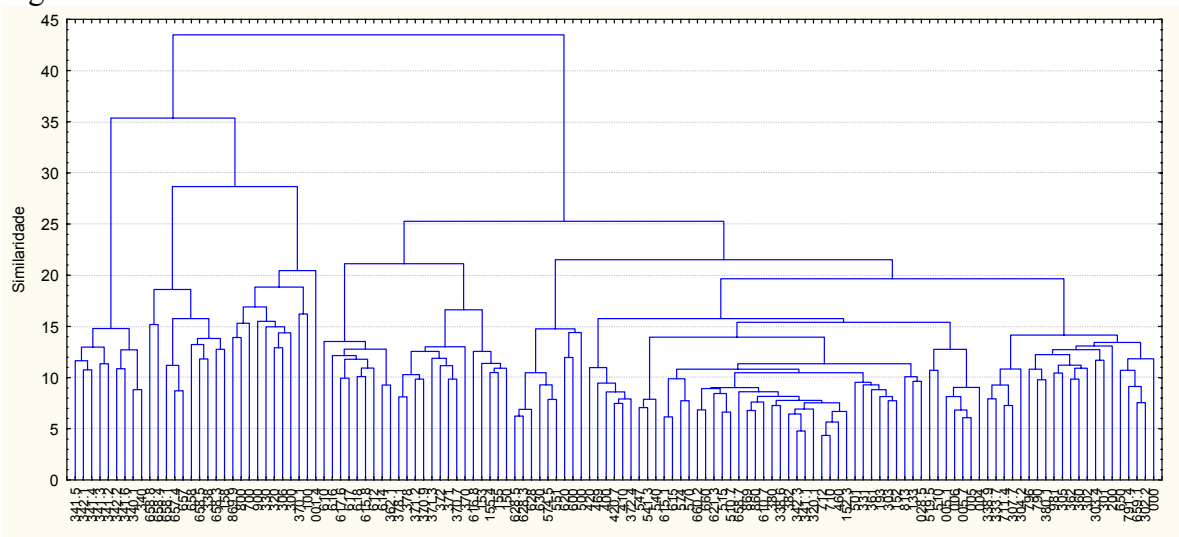


Figura 3. Dendograma com transações dos usuários por AS

Através da análise do dendograma foram gerados 34 grupos de grandes áreas como o apresentado na Tabela 1.

Tabela 1. Exemplo da tabela de grupos de grandes áreas de interesse

Grupo	Descrição do Grupo	AS	Descrição do AS
1	Direito	341.5	Direito penal
		342.1	Direito civil
		341.2	Direto constitucional
		342.2	Direito comercial
		341.6	Direito do trabalho
		340.1	Filosofia do Direito
		340	Direito

5.7.2 Classificação do acervo em grandes áreas

A classificação é uma tarefa muito utilizada em *data mining*. Consiste em examinar os aspectos de um objeto e atribuí-lo a um dos conjuntos de classes existentes. Assim,

classificam-se as obras do acervo da biblioteca em grandes áreas do conhecimento segundo a tabela gerada através da analise de *cluster* apresentada anteriormente.

5.7.3 Descrição do perfil dos usuários

Segundo Harrison (1998 p.181) "ás vezes o propósito de executar *data mining* é simplesmente descrever o que está acontecendo em um banco de dados complicado de maneira a aumentar o conhecimento das pessoas, produtos ou dos processos que produziram os dados".

A descrição do comportamento do usuário da biblioteca, através da análise de suas transações, objetiva identificar seu perfil de utilização de obras, podendo interagir com o mesmo através dos sistemas de SRI e DSI de forma personalizada.

No estudo do perfil, o primeiro nível de descrição seria a maior grande área de interesse, assim determinando a grande área de interesse do usuário. O próximo nível de descrição seria formado por três subáreas de interesse, no quarto nível da CDD, identificados através de uma análise das três principais áreas de transações do usuário. Tomamos como exemplo as transações de um usuário segundo as grandes áreas (Figura 3) e segundo CDD (Figura 4).

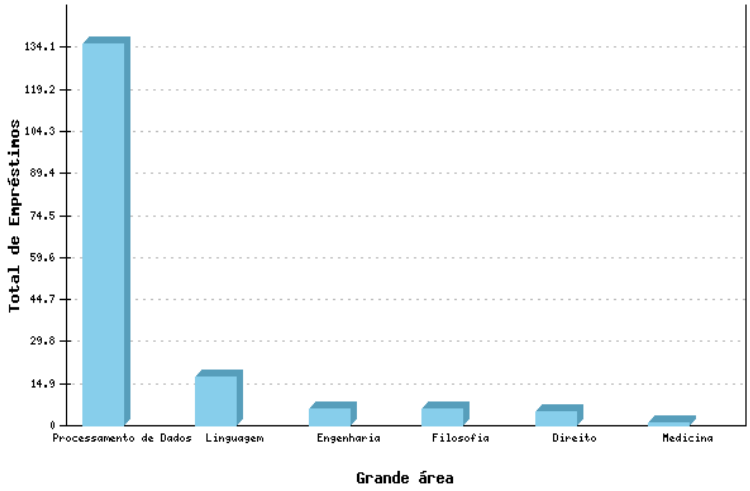


Gráfico 1. Transações usuários por grandes áreas

Pode-se verificar no gráfico 1 que o primeiro nível de descrição apresenta "processamento de dados" como a grande área de interesse do usuário em estudo.

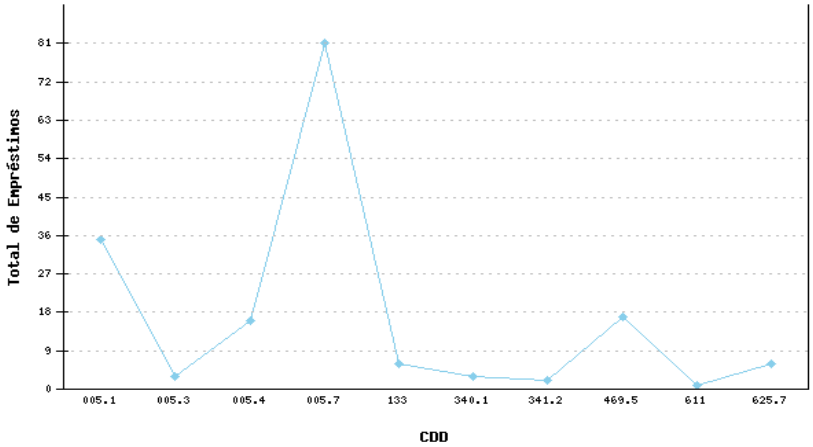


Gráfico 2. Transações usuários por CDD nível quatro

Como pode ser observado no gráfico 2, o segundo nível de descrição apresenta as três principais subáreas de interesse do usuário que são: 005.1, 005.7 e 469.5.

5.8. Plano de ação

Depois de identificado o perfil do usuário, torna-se possível personalizar os sistemas de recuperação e disseminação de informações. Para tanto, utiliza-se um sistema WEB (de SRI e DSI) que foi desenvolvido e personalizado dinamicamente ao perfil de cada usuário.

O sistema desenvolvido fica esperando requisições do servidor WEB quando a recebe processa e retorna páginas HTML com o conteúdo ao usuário personalizado. A arquitetura do sistema desenvolvido pode ser visto na Figura 4.

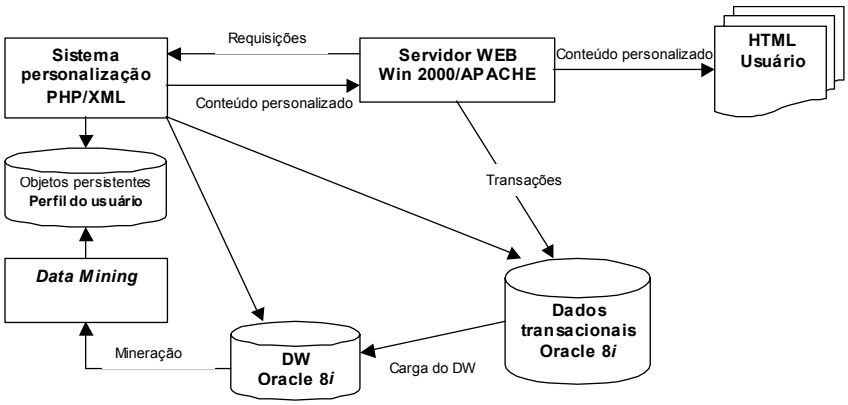


Figura 4. Arquitetura do sistema de personalização

O sistema conta com um banco de dados onde estão contidos os dados transacionais sobre as obras, usuários e suas transações, com um *data warehouse* onde estão os dados que serviram de fonte para a aplicação do *data mining*, e um objeto persistente o qual recebe os dados sobre o perfil do usuário. Quando o usuário faz uma requisição ao servidor WEB este recebe e a repassa para o sistema de personalização que recebe a requisição processa e envia a resposta de volta ao usuário personalizada.

5.9. Sistema WEB

Ao entrar no sistema é apresentada a tela de *login* onde devem ser informados o código e senha do usuário na biblioteca, após validação são carregados os dados do perfil do usuário para uma sessão no servidor, que funciona como objeto persistente ficando ativo até que o usuário saia do sistema.

A tela principal do sistema (Figura 5) é divida em três partes: menu superior, menu lateral, e corpo principal.

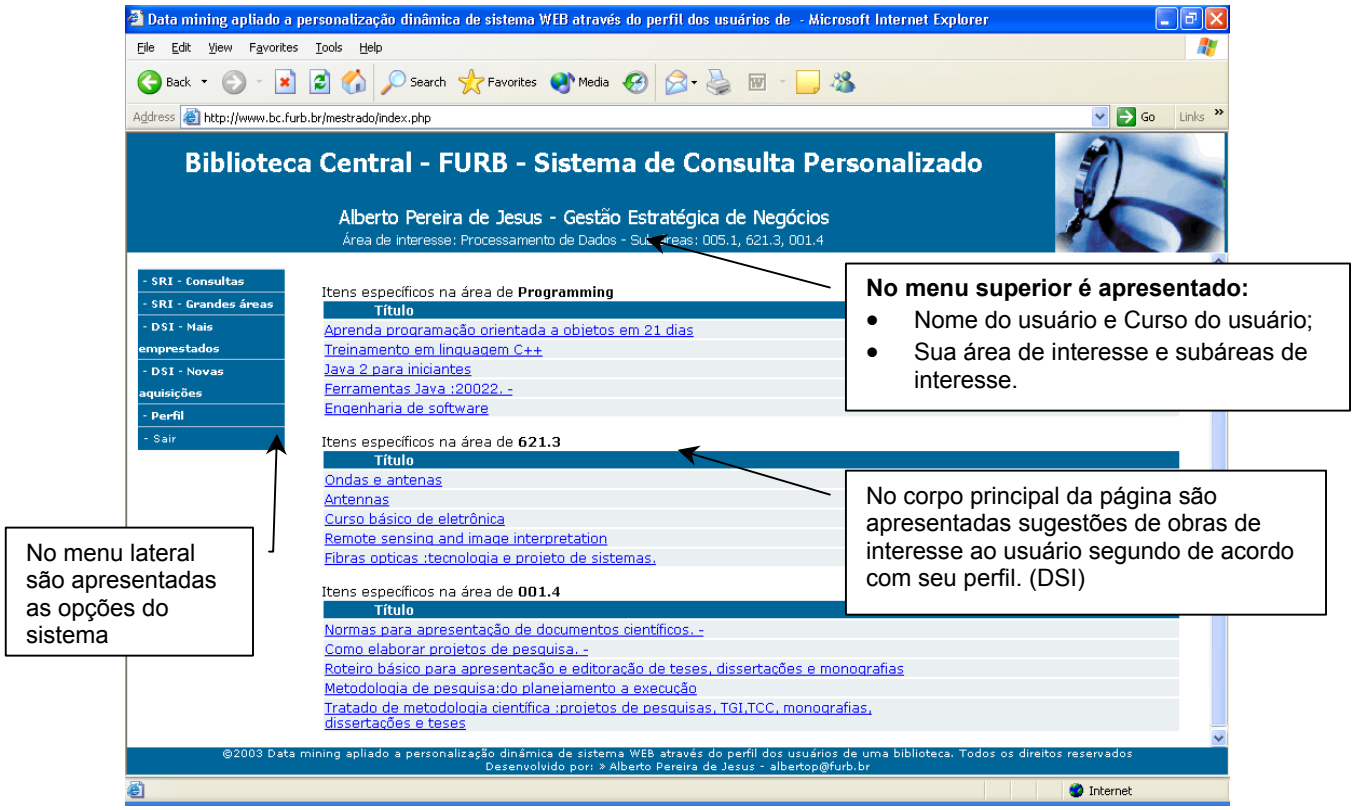


Figura 5. Tela principal do sistema

A tela de resultado da consulta (Figura 6) retornará os títulos encontrados no acervo segundo a expressão de busca determinada, ordenados conforme o perfil do usuário.

Resultado da consulta

Título	CDD	Relevância	Proximidade
Projeto & engenharia de software :teste de software	005.1	1	0
Qualidade & teste de software :engenharia de software, qualidade de software, qualidade de produtos de software, teste de software, formalização do processo de teste, aplicação prática dos testes	005.1	1	0
Guia completo ao teste de software	005.14	1	0.04
O teste gestáltico Bender para crianças	150.1982	999	145.0982
Tecnicas de exame psicologico e suas aplicacoes no Brasil :testes de aptidoes	155.28	999	150.18
Testes para admissão em empresas e empregos públicos	155.28	999	150.18

Figura 6. Tela resultado da consulta

5.10. Avaliação dos resultados

Através do modelo proposto e do protótipo desenvolvido foi possível melhorar o processo de recuperação e recomendações de obras através da identificação da relevância da mesma ao usuário.

6. Conclusão

Este trabalho propôs um modelo para extração automática do conhecimento sobre o perfil de usuários em bibliotecas. O modelo desenvolvido usa técnicas de *cluster*, classificação e descrição, fáceis de serem implementadas e interpretadas.

Os objetivos do trabalho foram alcançados. O *data warehouse* foi desenvolvido, o perfil dos usuários foi identificado com a aplicação de técnicas de *data mining* o sistema proposto implementado.

Quanto à tecnologia envolvida, acredita-se que está apenas nascendo e passará a fazer parte do nosso dia-a-dia. O mercado está em ampla expansão e com possibilidades de grandes negócios, pois a maioria das empresas possui grandes bancos de dados sem nenhuma utilização dos mesmos para tomada de decisões.

7. Referências

ARANHA, Francisco. **Análise de redes em procedimentos de cooperação indireta**: utilização no sistema de recomendações da Biblioteca Karl A. Boedecker. São Paulo: EAESP/FGV/NPP, 2000. 71p.

BARTOLOMEU, Tereza Angélica. **Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento**. 2002. 302f. Tese (Doutorado em Engenharia de Produção) – EPS. Universidade Federal de Santa Catarina, Florianópolis, 2002.

BERRY, Michael J. A, LINOFF, Gordon. **Data mining techniques**: for marketing, sales, and customer support. New York : J. Wiley E Sons, 1997. 454 p.

CARDOSO, Olinda Nogueira. Paes. Recuperação de Informação. **INFOCOMP Revista de Computação da UFLA**, Lavras, v.1, 2000. Disponível em: <http://www.comp.ufla.br/infocomp/e-docs/a2v1/olinda.pdf> Acesso em: 23 out. 2003.

DINIZ, Carlos Alberto R., LOUZADA NETO, Francisco. **Data mining**: uma introdução. São Paulo: ABE, 2000. 123p.

FUNARO, Vânia Martins B. O., CARVALHO, Telma de, RAMOS, Lúcia Maria S. V. Costa. **Inserindo a disseminação seletiva da informação na era eletrônica**. São Paulo: Serviço de Documentação Odontológica de Faculdade de Odontologia da USP. 17p.

HARRISON, T. H. **Intranet data warehouse**: ferramentas e técnicas para a utilização do data warehouse na intranet. Berkeley Brasil: São Paulo, 1998.

VELASQUEZ,Roberto M.G. et. al. Técnicas de Classificação para Caracterização da Curva de Carga de Empresas de Distribuição de Energia - Um Estudo Comparativo. **V Congresso Brasileiro de Redes Neurais**, 2001, Rio de Janeiro. Disponível em: <http://bioinfo.cpgei.cefetpr.br/anais/CBRN2001/5cbrn-6ern/artigos-5cbrn/>.

A Declarative Approach for Information Visualization

Adriane Kaori Oshiro
Andrea Rodrigues de Andrade
Maria da Graça Pimentel

Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Av. Trabalhador São-Carlense, 400 – Caixa Postal 668
13560-970, São Carlos, SP

{kaori, aandrade, mgp}@icmc.usp.br

Abstract. *Information Visualization investigates the use of visual and interactive information representations with the aim of reducing users' cognitive overhead as they analyze information. The objective of this work is to investigate mechanisms that allow presenting a large amount of information in Web-based platforms. We have built the iVIEW infrastructure that: (a) defines a declarative language based on XML Schema specifying a SVG-based visualization layout for information contained in XML documents; and (b) describes algorithms that execute the necessary steps to obtain a graphic representation of the information, implemented using XSLT. In this paper we describe the information elements and the visualization structures of iVIEW along with user-interaction resources. We also show the necessary steps for obtaining a graphic representation of the information using SVG.*

Keywords: Information visualization, XML-based language definition and processing, SVG.

1. Introduction

The recent growth of the Internet as a way to obtain information in the context of several application domains has demanded the incorporation of visual techniques that aid users in the task of interacting with this vast universe of information in an efficient and intuitive manner. Information Visualization has emerged as a research area that investigates the use of visual and interactive information representations with the aim of reducing users' cognitive overhead as they analyze information [Card et al., 1999].

The Web has become a repository for publishing applications such as newspapers and magazines, project-related documentation and educational material in general. One important feature in such publishing applications is that all information ever published is usually made available for users. As a result, the amount of information that a user has available to review grows as the time passes. In the case of a newspaper, for instance,

a user may be allowed to review an issue as old as needed. This is a typical application for visualization tools: to allow users to access and have some understanding of a huge amount of information. Moreover, it is important to offer users with not only a way to view all the information but also filter how the information is to be presented according to some specific attributes.

As far a typical publishing application is concerned, it is important to consider that each issue of a newspaper, for instance, is intrinsically related to others that have been previously published: this reflects the constant evolution of the contents of the published material. In such contexts, it is important that the visualization of the information of all issues be able to express the intrinsic relationship existent among the separate issues. Most visualization tools allow users to visualize existing relationships among the information items(e.g. [Hibino and Rundensteiner, 1996]), but they are associated to data with a specific structure and particular to specific domains [Polys, 2003].

SVG (Scalable Vector Graphics) is an XML based markup language for the specification of vector graphics, such as circles and polygons [W3C, 2003]. SVG documents can be visualized as a *standalone* document or embedded in HTML documents through the use of browser *plugins*. Several applications that use SVG can be found in the Web. For instance, Southard proposes the use of SVG to represent the structure of HTML and XML documents through an interactive SVG tree: branches correspond to document elements and the leaves correspond to attributes of these elements; by positioning the mouse cursor over a branch or a leaf, the name of an element or the content of its attribute is presented [Southard, 2001]. Examples by Adobe include graphs and images that simulate 3-D representations of molecules and interactive presentations of buildings and theaters [Adobe, 2002].

Applications that use SVG can present data from external sources, such as XML documents or Relational Databases [North et al., 2002]. For example, XML documents can be processed by algorithms contained in XSLT document in order to extract pertinent information towards generating SVG documents. This process was exploited in the implementation of a flexible and domain-neutral infrastructure, called iVIEW.

The main goal of this work is the investigation of mechanisms that exploit the processing of structured documents to allow the visualization of a large amount of information on the Web at the same time that supports a degree of interaction for presenting intrinsic relationships. We present iVIEW, that provides a mechanism for the visualization of information by means of automatically processing XML-based structured documents towards generating interactive SVG representations. Supported by a declarative language specified in XML Schema, iVIEW is extensible and independent of application domain by the use of XML format for data input.

In Section 2. we describe the iVIEW infrastructure while Section 3. details its use in the context of an XML publishing framework. Section 4. discuss our approach in the context of related work. Section 5. presents our conclusions and future work.

2. The iView infrastructure

The main goal of the iVIEW infrastructure is to provide the visualization of information extracted from XML documents using graphic representations in SVG. In order to achieve that goal, we have defined: a) a declarative language specifying a visualization layout for information (*layout.xml* in Figure 1); and b) algorithms to execute the necessary processing steps of documents to obtain a graphic representation of the information.

Processing steps. The processing of documents in iVIEW occurs in two stages. The first stage allows an intermediary transformation format to be obtained for a specific application (*Step 1a* in Figure 1). In the second stage, the intermediary specification is processed towards generating a final presentation specification (*Step 2* in Figure 1). It is this two-stage processing that gives generality to the overall transformation. Because the target documents of the overall processing are SVG, for graphics, and JavaScript, for user-interaction, the first stage also generates an interaction document (*Step 1b* in Figure 1).

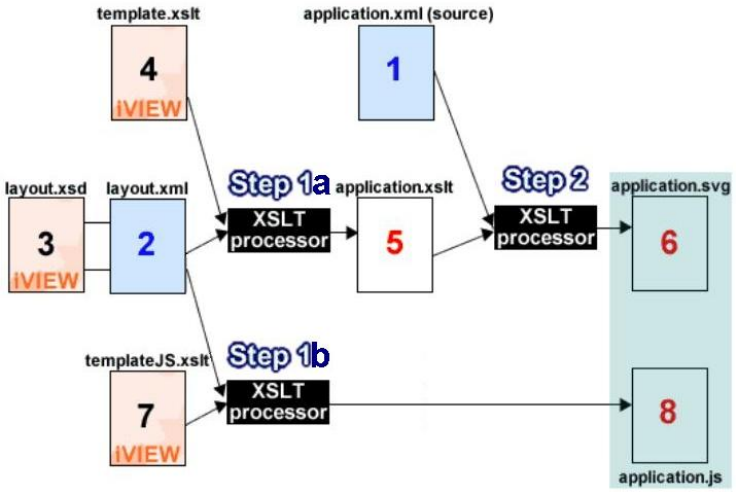


Figure 1: Documents, resources and processing steps of iVIEW for a graphic representation of information in SVG.

The numbered rectangles in Figure 1 represent input and output documents that are processed during each step. The developer is responsible for creating instances of *application.xml* (1), which correspond to the source of information that is to be visualized. The developer also designs *layout.xml* (2) that defines the structure of the presentation of the information, according to *layout.xsd* (3). Given these both input documents, the developer receives: *application.svg* (6) that contains a graphic representation in SVG of *application.xml* (1); and *application.js* (8) that contains JavaScript functions for providing users' interaction with *application.svg* (6).

The *application.svg* (6) and *application.js* (8) documents are generated by means of the iVIEW two-stage processing. In *Step 1a*, the generic XSLT stylesheet *template.xslt* (4) generates the XSLT stylesheet *application.xslt* (5) according to the specifications con-

tained in *layout.xml* (2). In order to generate *application.js* (8), *layout.xml* (2) is processed by means of the fixed XSLT stylesheet *templateJS.xslt* (7) in *Step 1b*. The *application.xslt* (5) stylesheet is specific for *application.xml* (1) to generate *application.svg* (6) in *Step 2*.

Visualization structure and groups of elements of iVIEW. In order to provide the visualization of information extracted from XML documents we have divided the basic visualization structure of iVIEW into two main areas: a) an area to present visualization elements groups with their attributes and relationships; and b) an area containing selection filters to show or hide determined elements. The actual positioning and dimensions of those areas are defined by the developer. Figure 2 are two examples of how the iVIEW visualization structure can be configured.

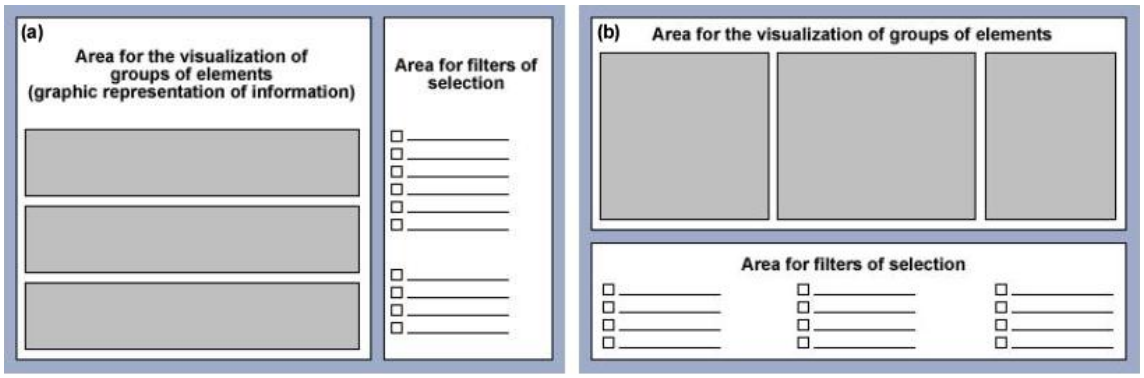


Figure 2: Example of iVIEW basic visualization structure configuration. a) Vertical display of main areas. b) Horizontal display of main areas.

We have defined a group of elements of visualization as a set of XML elements with same characteristics and structure, same attributes and that execute the same actions as response to users' interaction. For instance, the teaching staff of a given university can be considered as a group, where each instructor is associated to a visualization element. Assuming that all elements have the attributes "name", "department" and "email", we can say that they have the same nature and same characteristics. Figure 3 presents a possible visualization of three distinct elements groups:

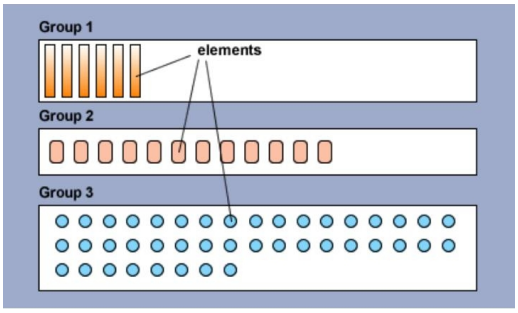


Figure 3: Distinct groups of elements.

As we can see in Figure 3, the elements are represented graphically by SVG shapes, such as bars, rectangles and circles. Each element represents an item of information, which contains one or more attributes that can be related to other elements. An important feature of groups of elements is the fact that all elements can be related to each other, in comparison to specific characteristics. This relationship can be visualized through filters and interaction functions.

Eventually, the amount of information items presented to users can be large and there are cases in which users may be interested in visualizing only elements with attributes containing a determined value. In order to provide the option of visualizing only determined elements, we have implemented elements selection filters. In this case, elements common attributes allow the visualization of intrinsic references among them.

We can also consider the visualization elements groups as structured lists. Therefore, the use and the combination of selection filters over these element groups can simulate the effect of union or intersection of lists. In Figure 4 we can observe an example of a selection filter. In Figure 4.a we can see a students group and a filter that select these students according to their graduation course. In Figure 4.b the filter is being used, and only students from ”Computer Science” and ”Physics” courses are visualized.

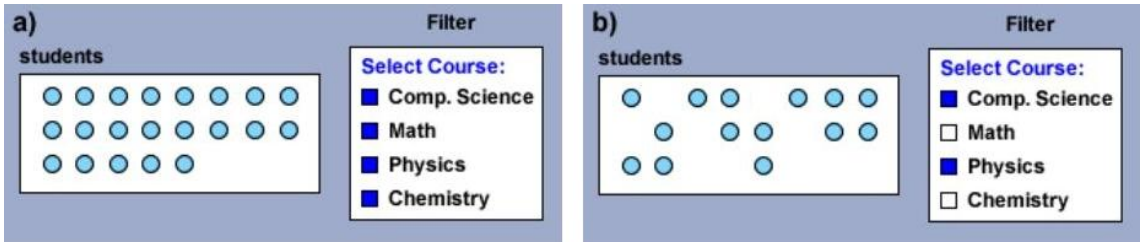


Figure 4: Elements selection filter. a) All students from all courses are shown. b) Only students from ”Comp. Science” and ”Physics” are shown.

Layout language. The goal of the first step for obtaining a graphic and interactive representation of an information set is the generation of a specific XSLT stylesheet (*application.xslt*) containing instructions that will generate a SVG representation of the associated information. This specific XSLT document will parse data to be visualized from the document *application.xml*.

An XSLT processor is used to generate the document *application.xslt*. As input, the developer must create the *layout.xml* document, based on the layout language defined in *layout.xsd*, that contains: (a) a specification of the general layout to be shown; (b) the layout of elements of information that will be visualized; and (c) some parameters necessary to establish the relationship between the elements.

Basically, the document *layout.xsd* defines the three main parts that establish the aspects and the properties of the final visualization: a) the layout, that makes possible the configuration of dimensions, positioning and colors of the main window, as well as

the SVG shapes that will represent information elements; b) the elements, which contains information of how the visualization elements are shown and what interaction functions can act over them; c) the filters, which contains information about the appearance and behavior of filters that are used.

As an example, the following code is a portion of a document *layout.xml* that refers to data about captured sessions in classrooms, references, students and instructors of a determined university:

```
<layout>
  <window  width="770" height="470" bgcolor="#648EB0"/>
  <title label="1o.Season 2001 - 2o.Season 2002"
    font_face="verdana" font_size="16" font_color="black"/>
  <script file="courses.js"/>
  <bars>
    <bar name="captured_sessions" stroke="3" height="30"/>
    <bar name="references" stroke="4" height="32"/>
  </bars>
  <circles>
    <circle name="instructors" radius="4" stroke_width="1"/>
    <circle name="students" radius="4" stroke_width="1"/>
  </circles>
</layout>
```

In the example, the attributes of the *window* element define the width and height of the final visualization area, as well as its background color. The *script* element contains the attribute file that defines which JavaScript document will be used to provide user interaction with the final visualization (the JavaScript document *application.js* is automatically generated). Finally, the *bar* and *circle* elements define how the visualization elements will be represented: captured sessions and references will be represented by bars; students and instructors will be represented by circles. Eventually, this representation could use any other basic shape of SVG, such as rectangles or other polygons.

In another part of *layout.xml*, the elements portion establishes the layout and elements properties that will be visualized:

```
<elements>
  <element path="courses/references" groupid="refs">
    <area x="10" y="320" width="545" height="40"
      stroke_width="2" stroke_color="white" bgcolor="#C6D5E1"/>
    <title name="references" x="10" y="315" font_color="white"
      font_size="11"/>
    <representation name="references" stroke_color="black"
      color="white"/>
    <initial_position min="1" max="50" x="18" y="324"/>
    <shift>10</shift>
    <attributes>
      <attribute name="title"/>
      <attribute name="author"/>
      <attribute name="course"/>
    </attributes>
  </element>
</elements>
```

The *element* element refers to information elements groups contained in the XML document *application.xml* — for instance, captured sessions, references, instructors and

students — and contains the definition of how these elements will be presented. The attribute path shows the hierarchical location of elements within the XML document *application.xml* that, in this example, are the elements containing data about references.

The attribute *groupid* contains an identifier to this group of elements and will be used subsequently by interaction functions with users. The *area* element defines the dimensions and properties of the space occupied by these elements in the final visualization. The *representation* element associates elements to their SVG representation defined previously in the general layout part (in the case of references, it was defined that they would be represented by a bar named ”references”). The *initial_position* and *shift* elements indicates, respectively, where the first element must appear and the space in pixels between each element. Finally, the attributes that will be visualized are defined by the *attributes* element.

The following code is the third and last part of the *layout.xml* document that is being presented as an example in this section, where the properties of the filters of selection of elements are defined:

```
<filters present="yes">
  <title label="Filters" font_size="14" font_color="blue"
    x="645" y="25"/>
  <area x="565" y="10" width="195" height="450" stroke_width="2"
    stroke_color="white" bgcolor="#98B4CB"/>
  <filter name="references" descriiption="Hide/show references:"
    font_color="blue" font_size="13" x="550" y="0">
    <groupelem groupid="refs" attributeid="2"/>
    <item label="Hypermedia" color="black" size="12"/>
    <item label="Multimedia" color="black" size="12"/>
    <item label="HCI" color="black" size="12"/>
    <item label="OS" color="black" size="12"/>
  </filter>
</filters>
```

The attribute *present* of the *filters* element indicates the presence of filters in the final visualization, since they may not be present if the developer defines so. The *title* and *area* elements define, respectively, the title and the area occupied by selection filters. Each filter used in the developer application is defined towards their properties. In this example, the filter named ”references” will act over the elements group which attribute *groupid* is ”refs”. The selection is composed by item which attributes label contain the values ”Hypermedia”, ”Multimedia”, ”HCI” and ”OS”. Therefore, users can visualize only the references about the course ”Hypermedia”, or about the courses ”Multimedia”, ”HCI” and ”OS”, or any combinations they prefer.

In order to guarantee that the *layout.xml* document is valid, it is necessary to follow the definitions contained in the *layout.xsd* document. This XML Schema defines the sequence and the possible number of occurrences of each element within the correspondent XML instance, as well as the data type that each element and attribute can have.

After the *layout.xml* document is properly specified, it will be processed by an XSLT processor together with the generic XSLT document *template.xslt*. The *template.xslt* document has the function of using the specifications contained in the *layout.xml* document

to generate the *application.xslt* document which, in turn, will process the *application.xml* document. As a result, a SVG document (*application.svg*) is produced and it presents the information of the document *application.xml* in the format specified by the developer.

The portions of documents presented in this subsection refer to an application for visualization of data about captured sessions in classrooms, references, students and instructors of a university. The XML document for this application (*application.xml*) is relatively large and information in it cannot be visualized in a single screen. However, the visualization of the information in a single screen can be possible by means of the transformation of the XML document into SVG.

Figure 5 is the SVG representation of the application in its complete form, where the information elements are represented by vertical bars (captured sessions and references) and circles (instructors and students), according to parameters defined by the developer in the first step. Figure 5 also shows the execution of generated interaction functions obtained in the last step.

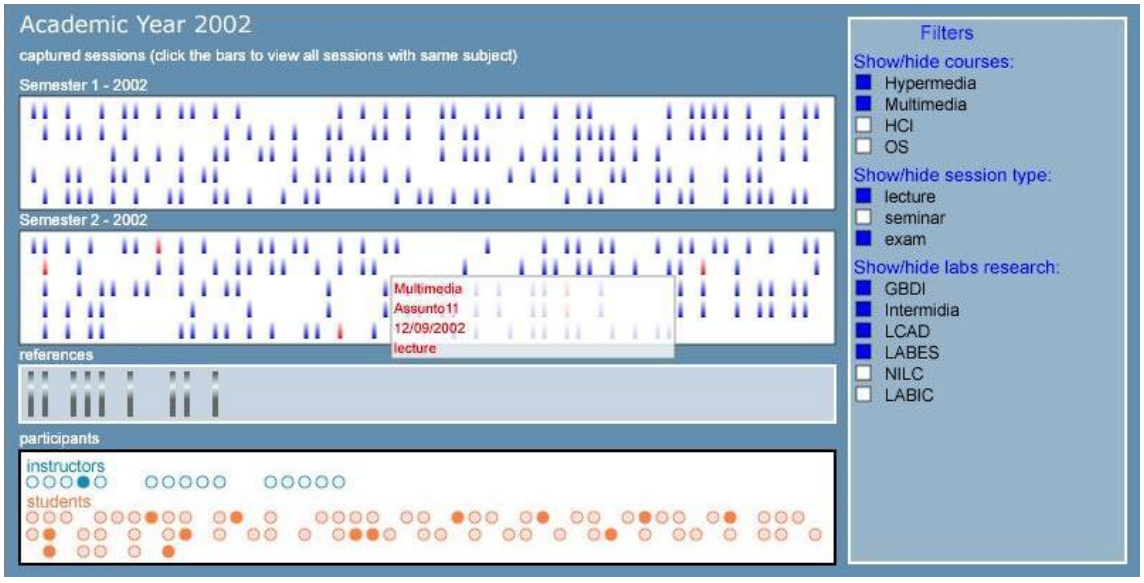


Figure 5: SVG representation associated to a JavaScript document with functions that provide interaction with the elements displayed.

3. iVIEW in use

As the iVIEW infrastructure is XML-based, its processing steps could be implemented in an XML publishing framework. We exploit the Cocoon Java Framework, developed by the Apache Software Foundation, that supports Web publishing based on the processing of XML documents [Apache, 2002].

The processing in Cocoon is pipeline-based: as a result of a user request for a document, XML documents enters the pipeline, are processed and passed by means of

SAX events to the next processor in the pipeline, until they exit the pipeline in a format that can be delivered over the Web. The following is a pipeline defined according to iVIEW processing steps[1]:

```
<map:pipeline>
  <!-- Step 1a -->
  <map:match pattern="application.xslt">
    <map:generate src="layout.xml"/>
    <map:transform src="template.xslt"/>
    <map:serialize type="xml"/>
  </map:match>
  <!-- Step 1b -->
  <map:match pattern="application.js">
    <map:generate src="layout.xml"/>
    <map:transform src="templateJS.xslt"/>
    <map:serialize type="text"/>
  </map:match>
  <!-- Step 2 -->
  <map:match pattern="application.svg">
    <map:generate src="application.xml"/>
    <map:transform src="cocoon:/application.xslt"/>
    <map:serialize type="svg"/>
  </map:match>
</map:pipeline>
```

4. Related work

Several work in the context of information visualization exploit the use of XML based languages to store and transform data. One example is an application that uses XML and XSLT to manipulate video information [Christel et al., 2001]. This application used Java applets to display information obtained as result of queries. Our implementation is based on the specification of XML documents that are processed by general XML processors.

Other example is an application that provides the visualization of documents that embed contextual, data-driven information components using SVG [Weber et al., 2002]. However, the document generation is manual or semi-automatic; one of the main efforts of our implementation was to provide automatic generation of SVG representations and other documents related user-interaction.

XML-based languages are also exploited by an application that uses X3D and VRML to visualize data stored in CML (Chemical Markup Language), which is an XML-based language for representing chemical data [Polys, 2003]. XSLT is used to transform information into X3D and VRML. In the same context, data stored in CML has also been represented by interactive graphics in SVG [Adobe, 2002]. Our implementation is more generic in terms of information domain, since any XML specification can be used. Moreover, our work supports the use of any SVG-based graphic representation.

[1]Example available from http://iclass.icmc.usp.br/iview

5. Conclusions and further work

To Web users, the task of finding, interpreting and interacting with a vast universe of information is a non-trivial task. Information Visualization is a research area with the objective to provide interactive ways to represent data.

Supported by a declarative language specified in XML Schema, the iVIEW infrastructure allows the developers to define graphic SVG information representations so that users can view a great amount of elements in a single visualization in an interactive way.

Compared to current efforts, our implementation is (a) more general in terms of the format of the input data specified in XML documents; (b) does not require the developer to specify XSLT transformations, JavaScript functions or SVG representations; (c) allows reuse of presentation templates as layout documents; (d) allows the identification of relationships among items of information towards the visualization of the overall information.

Possibilities to extend the resources offered by the current infrastructure include: to investigate ways to allow the establishment of information relationship after the generation of the SVG representation; to investigate other resources that SVG is capable to provide so we can add new interaction functions; to create graphic publishers to define the visualization layout.

References

Adobe (2002). SGV Zone Demos. http://www.adobe.com/svg/demos/main.html.

Apache (2002). The Apache Software Foundation: Apache Cocoon 2.0. URL:http://cocoon.apache.org/2.0/.

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers.

Christel, M. G., Maher, B., and Begun, A. (2001). XSLT for tailored access to a digital video library. In *1st. ACM/IEEE-CS joint conference on Digital libraries*, pages 290–299.

Hibino, S. and Rundensteiner, E. A. (1996). MMVIS: A MultiMedia Visual Information Seeking Environment for Video Analysis. In *Proc. ACM Multimedia'96 Conference*, pages 15–16.

North, C., Conklin, N., and Saini, V. (2002). Visualization Schemas for Flexible Information Visualization. In *Proc. IEEE InfoVis 2002 Symposium*, pages 15–22.

Polys, N. F. (2003). Stylesheet transformations for interactive visualization: towards a Web3D chemistry curricula. In *Proc. 8th. International Conf. on 3D web technology*, pages 85–90.

Southard, J. (2001). XML Grove. http://www.jeffsouthard.com/demos/grove/.

W3C (2003). World Wide Web Consortium, Scalable Vector Graphics (SVG) 1.1 Recommendation. http://www.w3.org/TR/SVG11.

Weber, A., Kienle, H. M., and Müller, H. A. (2002). Live documents with contextual, data-driven information components. In *Proc. 20th annual international conference on Computer documentation*, pages 236–247.

Um projeto de metodologia para escolha automática de descritores para textos digitalizados utilizando sintagmas nominais

Renato Rocha Souza[1], Lidia Alvarenga[1]

[1]Escola de Ciência da Informação – Universidade Federal de Minas Gerais (UFMG)
Avenida Antônio Carlos, 6627 31270-010 Belo Horizonte, MG – Brasil

{rsouza,lidiaalvarenga}@eci.ufmg.br

Abstract. *It can be noticed that the indexing and representation strategies nowadays seems to be near the exhaustion, and it is worth to investigate new approaches to the indexing and information retrieving systems. Among these, a branch tries to consider the intrinsic semantics of the textual documents using noun phrases as descriptors instead of single keywords. We present in this article a methodology that is being developed in the scope of a doctorate research.*

Resumo. *Com o aparente esgotamento das estratégias atuais de representação e indexação de documentos, faz-se necessário investigar novas abordagens para sistemas de recuperação de informações. Dentre estas abordagens, há uma vertente que busca levar em conta a semântica intrínseca aos documentos textuais, e uma das formas de fazê-lo é através da utilização de sintagmas nominais como descritores, ao invés de palavras-chave. Uma metodologia para atingir tal propósito, que está sendo desenvolvida no escopo de uma tese de doutorado, é apresentada neste artigo.*

1. Introdução

Para lidar com os constantes e ininterruptos ciclos de criação e demanda de informação, há muito vêm sendo criados sistemas de recuperação de informações[1] que utilizam diversas tecnologias mecânicas e digitais de computação, para gerenciar grandes acervos de documentos. Podemos citar, dentre eles, a Internet, as intranets empresariais com seus portais corporativos, e as bibliotecas digitais.

Neste contexto, este artigo apresenta uma pesquisa em andamento, desenvolvida no âmbito do curso de doutorado do autor, no Programa de Pós Graduação em Ciência da Informação da Universidade Federal de Minas Gerais. A pesquisa pretende contribuir para enfrentar alguns dos muitos desafios que surgem quando lidamos com massivas quantidades de dados, como nos grandes acervos de documentos digitais,

[1] Entende-se, no escopo deste trabalho, que os sistemas de recuperação de informações são sistemas, usualmente baseados em tecnologias digitais, que lidam com a organização e o acesso aos itens de informação, desempenhando as atividades de representação, armazenamento e recuperação destes itens.

notadamente quando estes precisam ser regularmente organizados e pesquisados, recuperando em tempo hábil informação relevante para algum objetivo específico.

Com o aparente esgotamento[2] das estratégias tradicionais de busca em sistemas de recuperação de informações, entendemos que a melhoria da eficácia do serviço ao usuário dos sistemas depende dos resultados em diversas linhas de pesquisa, em todo o espectro da cadeia de processos de tratamento da informação. Temos como hipótese de trabalho que as principais frentes de atuação são as seguintes:

1. A exploração das informações semânticas e semióticas intrínsecas aos dados, de forma a expandir a compreensão das unidades e padrões de significado em textos, imagens e outras mídias;
2. O desenvolvimento de novas possibilidades de marcação semântica dos dados utilizando-se metalinguagens, criando espécies de índices acoplados aos próprios documentos com termos amplamente consensuais e não ambíguos, para que estes possam ser mais facilmente manipulados e identificados por computadores e outros dispositivos e, como conseqüência, pelos usuários;
3. O desenvolvimento de estratégias de apresentação da informação recuperada nas buscas sob formas altamente significativas, ou contextuais[3] - como em algumas interfaces gráficas – de forma que as relações entre os conceitos, e em conseqüência, os contextos, sejam evidentes; e também por estratégias que busquem estimular os vários órgãos sensoriais ao mesmo tempo – como nas ferramentas multimídias – para que a absorção das informações pelos usuários seja maior. Através destas interfaces e estratégias, as informações podem ser apresentadas de forma a possuírem conexões visuais aos seus contextos de origem, permitindo ao usuário refinar os resultados através da definição das conexões pertinentes e a exclusão das conexões geradas pelo ruído informacional;
4. A construção e manutenção de perfis personalizados de utilização, de forma que o SRI "aprenda" com a forma de trabalho do usuário e possa utilizar estas informações específicas para melhorar a estratégia de busca do SRI.

Uma abordagem completa para a organização e a recuperação de informações, visando a melhoria dos Sistemas de Recuperação atuais, deve unir estas estratégias e soluções, buscando:

- A indexação dos documentos utilizando representações mais significativas, de modo a aumentar e melhorar os pontos de acesso e a relevância das informações recuperadas;
- Prover uma forma adequada de apresentar as informações recuperadas aos usuários, de maneira que sejam intuitivas e facilmente compreensíveis;

---

[2] As estratégias tradicionais baseiam-se em modelagens dos documentos a partir da distribuição de suas palavras-chave. Embora existam propostas de avanços, parece haver um limite para a eficácia de tais estratégias.

[3] Informação apresentada sem desprezo do contexto que lhe confere sentido.

- Utilizar no processo de indexação padrões universais de registros de metadados para que os sistemas sejam interoperáveis entre si;
- Adaptar-se continuamente ao usuário, sendo preferível que possa aprender com a forma com que trabalha, de modo que as buscas sejam continuamente refinadas através de um trabalho de personalização.

Existem hoje diversas tentativas, mais ou menos coordenadas, de se abordar estas ações fundamentais, mas uma real integração demandaria a pesquisa em diferentes áreas do conhecimento e campos de pesquisa, como a ciência da informação, a lingüística, a ciência da computação, a sociologia, a antropologia, a comunicação, a psicologia cognitiva, entre outras.

De maneira isolada, há pesquisas em cada uma destas vertentes, mas é pouco explorada a utilização da semântica embutida nos próprios documentos, ou seja, das potencialidades intra-textuais da linguagem natural, para automatizar e melhorar as tarefas de indexação, organização e recuperação de informações.

Pesquisas nesta área incluem o uso de estruturas profundas da linguagem natural, como os sintagmas verbais e nominais, para indexação e recuperação [KURAMOTO, 1996 e 1999; MOREIRO et al, 2003]; e de ferramentas de representação de relacionamentos semânticos e conceituais, como os tesauros, para ampliar a gama de informações recuperadas e aferição de contextos [SPARCK JONES & WILLETT, 1997, pp. 15-20]; além de outras estratégias derivadas da lingüística e da ciência da informação. Todas estas estratégias são fortemente atreladas ao idioma, o que faz com que os possíveis resultados da pesquisa tenham uma aplicação circunscrita ao contexto da língua da comunidade em questão. As metodologias, entretanto, são generalizáveis e sua aplicabilidade a outras linguagens é perfeitamente possível.

Neste projeto, pretende-se apresentar uma metodologia para aproveitar o potencial de uso dos sintagmas nominais como descritores de documentos em processos de indexação. Parte-se da hipótese de que os sintagmas nominais, pelo maior grau de informação semântica embutida, podem vir a se tornar mais eficazes do que as palavras-chave usualmente extraídas e utilizadas como descritores em outros processos automatizados de representação de documentos, tais como os observados nos mecanismos de busca da Internet, ou em sistemas de leitura das palavras-chave fornecidas pelo autor dos documentos.

Alguns trabalhos anteriores se apresentam como marcos a partir dos quais se pretende avançar; dentre eles, a pesquisa sobre a viabilidade do uso dos sintagmas nominais para sistemas de recuperação de informações de KURAMOTO [1996 e 1999], e as ferramentas para marcação sintática do português e automatização da extração de sintagmas nominais desenvolvidas no âmbito dos projetos da Southern Denmark University [BICK, 2000], de VIEIRA [2000] e do PROJETO DIRPI [2001]. A partir destes resultados e ferramentas, pretende-se propor uma metodologia de escolha automática de descritores para documentos que utilize os sintagmas nominais em vez de palavras-chave para documentos textuais digitalizados em língua portuguesa.

2. Sintagmas nominais e sistemas de recuperação de informações

Entendemos por **sintagmas** certos grupos de palavras que fazem parte de seqüências maiores na estrutura de um texto, mas que mostram um grau de coesão entre eles

[PERINI, 1995]. Os constituintes ou sintagmas podem ou não ser facilmente identificáveis, sendo que por vezes é necessário recorrer a outros recursos para que seja feita a "demarcação" sintática. Perini acredita que a intuição "subjetiva, mas nem por isso duvidosa" que nos permite separar a oração em seus constituintes imediatos pode ser caracterizada através de critérios puramente formais [1985, pp. 42-43], mas há quem defenda que a identificação dos constituintes é somente completa através de uma abordagem cognitiva e amplamente contextual [LIBERATO, 1997], que só é esperada na análise do discurso[4] e na pragmática[5]; ou através de outros modelos gramaticais, como a análise transformacional [RUWET, 1975, pp.155-212 e 223-279]. Para a análise semântica, há também o problema das situações anafóricas, que ocorrem quando uma estrutura de uma oração se apresenta reduzida porque ocorre na vizinhança de outra estrutura oracional de certa forma paralela, dependendo desta para sua total compreensão [PERINI, 1986, p. 57].

De acordo com MIORELLI [2001], os sintagmas nominais podem ser entendidos – e tratados – de forma sintática, privilegiando a forma; ou semântica, buscando os significados maiores; cada uma com suas especificidades e implicações. A abordagem semântico-pragmática, utilizada por LIBERATO [1997], não prescinde de um "interpretador de contextos", natural na cognição humana, mas dificilmente implementado em heurísticas de inteligência artificial. A forma sintática, como analisada por PERINI [1986, 1995 e 1996] está mais relacionada à estrutura das orações em si, e é mais facilmente tratada computacionalmente. Assim como no trabalho de MIORELLI [2001], esta é a abordagem que será utilizada no âmbito deste projeto, da mesma forma que, provavelmente, em quaisquer abordagens, e com quaisquer ferramentas, que busquem a automatização de extração dos sintagmas nominais.

Sistemas de recuperação de informações usualmente adotam termos índices para indexação de documentos, sendo que estes termos índice são usualmente palavras-chave. Há uma idéia fundamental embutida de que a semântica dos documentos e das necessidades de informação do usuário podem ser expressas através destes conjuntos de palavras, o que é, claramente, uma grande simplificação do problema, porque grande parte da semântica do documento ou da requisição do usuário é perdida quando se substitui o texto completo por um conjunto de palavras [BAEZA-YATES & RIBEIRO-NETO, 1999, p.19].

Há, na literatura, registros de algumas tentativas de otimizar a organização dos documentos em SRIs através de um processamento aprofundado da linguagem natural dos documentos. Dentre elas, a identificação de "grupamentos de substantivos" (*noun groups*), ao invés de palavras-chave, se afigura uma boa estratégia para seleção de termos de indexação, uma vez que os substantivos costumam carregar a maior parte da semântica de um documento, ao invés de artigos, verbos, adjetivos, advérbios e conectivos. Esta proposta estabelece uma visão conceitual do documento [ZIVIANI, in BAEZA-YATES & RIBEIRO-NETO, 1999, pp.169-170]. Os grupamentos de substantivos são conjuntos de nomes nos quais a distância sintática no texto (medida

[4] Estuda a estrutura e a interpretação dos textos.

[5] Ocupa-se da relação dos enunciados lingüísticos com a situação extralingüística em que se inserem [PERINI, 1995].

pelo número de palavras entre dois substantivos) não excede um limite predefinido. Uma metodologia que extrapola esta proposta é a identificação dos sintagmas nominais e o seu uso como descritores, como proposto neste projeto.

SALTON & MCGILL [1983, pp. 90-94] discutem algumas abordagens teóricas para o uso de métodos lingüisticos na recuperação de informações ; entre elas, a análise da estrutura sintática (*parsing*) dos documentos de forma a identificar as estruturas sintagmáticas. Estes autores, entretanto, apontam as dificuldades intrínsecas ao processo de análise semântica através da análise sintática e exemplificam casos em que é impossível o reconhecimento não ambígüo de relações semânticas através dos componentes da sentença, sugerindo que um modelo baseado em gramáticas transformacionais poderia trazer melhores resultados. Neste ponto, parecem então concordar com LIBERATO [1997], que entende que a análise completa das estruturas semânticas só é possível através da análise cognitiva dos contextos. Ao indicar a maior eficácia relativa dos algoritmos de geração de frases baseadas em frequência de palavras, talvez apontem uma alternativa para a melhoria do algoritmo proposto neste trabalho. Outra alternativa apontada é a interferência humana no processo de desambiguação através de uma interface, o que seria pouco desejável num processo que pretende ser automático.

Um importante caminho de pesquisa que visa resolver os problemas de desambiguação semântica através da análise dos contextos é resolução de correferência, ou resolução anafórica [VIEIRA, 1998 e 2000; SANT'ANNA, 2000 ; ROSSI et al, 2001; GASPERIN et al, 2003]. Uma cadeia de correferência é uma seqüência de expressões em um discurso que se referem a uma mesma entidade, objeto ou evento. Essas cadeias são úteis para a representação semântica de um modelo de domínio, e podem melhorar a qualidade dos resultados em diversas aplicações de processamento de linguagem natural, como recuperação e extração de informações, geração automática de resumos, traduções automáticas, entre outros [ROSSI et al, 2001]. O processo de resolução de correferências envolve a identificação e extração dos sintagmas nominais.

LE GUERN e BOUCHÉ [apud KURAMOTO, 1999] apontam o sintagma nominal como a menor unidade de informação contida em um texto. O grupo de pesquisas SYDO, ao qual pertencem estes pesquisadores, tem como fundamento teórico a utilização de sintagmas nominais como descritores [Ibidem, 1996]. Ao trabalhar em parceria com este grupo, KURAMOTO [1999], em sua tese de doutorado, desenvolveu uma pesquisa fundamental para a consideração da utilização de sintagmas nominais como descritores. Já em um trabalho anterior, KURAMOTO [1996] vislumbrou a maquete proposta na tese e já apontava o potencial natural de organização dos sintagmas nominais que, se explorado convenientemente, poderia propiciar aos usuários maior facilidade no uso de um SRI e resultados mais precisos em resposta a um processo de busca de informação.

O sistema desenvolvido por Kuramoto pode ser considerado uma inspiração para o presente trabalho, na medida em que, em ambos, busca-se uma alternativa para uma melhor indexação utilizando-se sintagmas nominais. Entretanto, em sua maquete, "A extração dos sintagmas nominais foi realizada de forma manual, simulando uma extração automática. Este procedimento foi adotado em função da não-existência ainda de um sistema de extração automática de SN em acervos contendo documentos em língua portuguesa". [1996, p.6]. Ao menos um sistema deste tipo, entretanto, se

encontra hoje disponível, e foi disponibilizado para o propósito deste trabalho [GASPERIN et al, 2003]. Uma outra diferença fundamental é o objetivo. Se no projeto de Kuramoto buscava-se apresentar uma maquete de um SRI baseado em sintagmas nominais, o objetivo deste trabalho é propor uma metodologia de auxílio à indexação automática utilizando uma metodologia aplicada sobre os sintagmas nominais extraídos automaticamente. Diferenças a parte, o fundo filosófico é bastante comum.

3. O método proposto

Acreditamos que, neste ponto, todo o cabedal teórico necessário ao entendimento do contexto no qual se insere o projeto já tenha sido apresentado e possa ser corretamente entendido. Nesta altura, cabe apresentar a metodologia proposta para consecução dos objetivos.

A) Para o objetivo geral: "Propor uma metodologia para escolha semi-automática de descritores para documentos textuais digitalizados em língua portuguesa, utilizando as estruturas sintáticas e semânticas conhecidas como sintagmas nominais"; pretendemos perfazer os seguintes passos, em seguida explicitados e comentados:

1. Escolher um corpus considerável de textos publicados recentemente em meio eletrônico em revistas científicas da área de Ciência da Informação;
2. Analisar o corpus escolhido, retirar suas palavras-chave atribuídas pelos autores e informações adicionais de formatação, e extrair os sintagmas nominais do corpo do texto, utilizando as ferramentas detalhadas adiante;
3. Verificar a freqüência de incidência dos sintagmas nominais e adotar uma lógica para escolha dos mais significativos;

Neste ponto, talvez esteja uma das partes mais críticas da metodologia. A lógica para escolha dos sintagmas nominais relevantes está para ser estabelecida através da manipulação dos dados empíricos. Pode-se esperar, entretanto, que venha a ser derivada dos algoritmos de extração de palavras-chave baseados na lei de Zipf (freqüência simples com descarte dos picos, pesos relacionados à freqüência inversa nos documentos, valor discriminatório dos termos) apresentados anteriormente ou mesmo o algoritmo composto proposto por SALTON & MCGILL [1983, pp.71-75]. Há que se fazer as adaptações necessárias ao fato de não mais estarmos tratando de palavras-chave, mas sim de sintagmas nominais. Não são descartadas, entretanto, as metodologias de busca sequencial [BAEZA-YATES & RIBEIRO-NETO, 1999, pp. 209-215]. Espera-se que, após a obtenção de resultados satisfatórios em um pequeno subconjunto dos textos, o restante do corpus seja usado para validação da metodologia escolhida.

4. Verificar a incidência dos sintagmas nominais escolhidos em um tesauro na área da Ciência da Informação; separar os verificados em conjuntos doravante denominados: a) os que constam no tesauro e b) os que não constam no tesauro;
5. Comparar, separadamente, os sintagmas dos conjuntos a) e b) definidos acima com as palavras-chave escolhidas pelos autores dos textos e com o assunto do texto, na forma em que puder ser depreendido. Analisar os resultados;

6. Julgar, dentre aqueles que não constam do tesauro, quais deveriam constar; separar estes sintagmas nominais em conjuntos doravante denominados: c) sintagmas nominais que deveriam constar no tesauro; d) sintagmas nominais referentes a conceitos relevantes de áreas afins e; e) sintagmas nominais devem ser ignorados de forma semelhante às *stopwords*. Os sintagmas recolhidos em c) serão considerado para fins de validação dos próximos sintagmas, enquanto os sintagmas em d) serão analisados mais detidamente, pois a metodologia poderia ser ampliada com a utilização de tesauros de áreas correlatas. Os sintagmas em e) serão descartados das próximas análises;

A proposta metodológica deve sofrer alterações, na medida em que os dados empíricos forem manipulados e analisados. No entanto, este trabalho não teria sido possível sem as ferramentas de extração automática que, assim como os corpora, foram gentilmente cedidos pelos proprietários e desenvolvedores. Em seguida passamos à descrição destas ferramentas.

4. Ferramentas utilizadas

O trabalho de análise proposto na metodologia acima descrita é talvez a ponta do iceberg de todo o esforço computacional necessário, compreendido no processo. Para que seja possível a análise dos descritores, os sintagmas nominais tiveram que ser extraídos, no caso, automaticamente e de forma bastante veloz. Os textos dos corpora foram escolhidos pelo autor e transformados em formato de texto simples. Em seguida, foram submetidos sucessivamente ao processamento da ferramenta "Palavras" da Southern University of Denmark e o software "Palavras Extractor" desenvolvido em conjunto pela Universidade do Vale do Rio dos Sinos (Unisinos) de São Leopoldo e a Universidade de Évora, em Portugal. Os pesquisadores da Unisinos e da Universidade de Évora cederam, para os propósitos deste trabalho, uma interface integrada através da qual grande parte do processamento automático envolvido, inclusive o desempenhado pelo *site* dinamarquês, foi realizado, durante os meses de agosto e setembro de 2003. Em seguida vamos descrever em mais detalhes estas ferramentas.

4.1. O VISL e o processador "Palavras"

A Southern University of Denmark, desenvolveu e tornou público uma ferramenta de processamento morfo-sintático de textos digitalizados em português chamada "Palavras", que faz parte de um conjunto de ferramentas multilinguais chamado VISL (Virtual Interactive Sintax Learning), disponível no endereço da Internet: http://visl.sdu.dk/visl/. No VISL, várias ferramentas, para cada um dos idiomas suportados, operam em modo automático ou semi-automático, nos quais um usuário submete sentenças ou textos completos em uma das linguagens admitidas (dentre as quais, o português) e recebe de volta os textos marcados. As análises podem ser feitas em diferentes níveis (morfológico, sintático, semântico) e em várias formas de visualização, como textos simples, árvores sintáticas ou marcação com cores [BICK, 1996, 2001 e 2003]. O projeto VISL é altamente orientado a produtos e processos, uma vez que novas ferramentas tem sido constantemente disponibilizadas gratuitamente na Internet na medida em que os protótipos se mostram funcionais. O VISL é baseado em um emaranhado de páginas HTML, scripts CGI (commom gateway interface), Java e

PERL, e oferece uma interface gráfica que permite aos usuários uma diversidade de opções [VISL, 2003].

Uma das possibilidades de marcação oferecidas pelas ferramentas do *site* indica as categorias gramaticais e a função de cada palavra no contexto de uma oração. Através desta marcação e um processamento posterior, é possível extrair os sintagmas nominais das sentenças de um texto. Este pós-processamento pode ser feito manualmente, através da análise das funções e da proximidade das palavras, ou pode ser automatizado, o que é o objetivo da ferramenta "Palavras Extractor", descrita a seguir.

4.2. A extração automática de sintagmas nominais

A partir da ferramenta computacional "Palavras" do VISL, o Departamento de Lingüística Computacional Aplicada do Centro de Ciências Exatas e Tecnológicas da Universidade do Vale do Rio dos Sinos, sob a coordenação da professora doutora Renata Vieira, em parceria com o departamento de Informática da Universidade de Évora, de Portugal; desenvolveu, no escopo do projeto de cooperação DIRPI [PROJETO DIRPI, 2001], um conjunto de programas de interface e de pós-processamento dos resultados, chamados internamente de "Palavras Extractor". Os programas estabelecem um acesso ao *site* VISL, enviam textos para o analisador sintático PALAVRAS para o português [Bick, 2000 apud GASPERIN et al, 2003]. A saída desse analisador é convertida em um conjunto de arquivos XML: o arquivo de palavras (elementos <word>); um arquivo com as categorias morfo-sintáticas (POS - Part Of Speech) das palavras do corpus e um arquivo com as estruturas sintáticas das sentenças, representadas por "chunks" [GASPERIN et al, 2003]. A partir destes três arquivos XML gerados, pode-se trabalhar com mais facilidade e desenvoltura em comparação com o *output* do site VISL, pois através do uso de folhas de estilo (XSL) específicas, é possível então extrair os sintagmas nominais de qualquer texto ou corpus da língua portuguesa. Assim como são extraídos os sintagmas nominais, é possível extrair outras instâncias gramaticais, dependendo do interesse da pesquisa em questão, bastando para tanto o desenho de uma nova folha de estilo. Os sintagmas nominais utilizados neste projeto foram obtidos utilizando-se a folha de estilo específica para extração de sintagmas nominais, cedida pela pesquisadora da Unisinos Claudia Camerini Correa Perez, e o software XML SPY (http://www.altova.com), utilizado para aplicação da transformação XSL nos arquivos XML gerados. O resultado final são arquivos HTML contendo os sintagmas nominais na seqüência em que ocorrem no texto, desde os sintagmas nominais máximos até os sintagmas aninhados na estrutura máxima.

5. Conclusões

Ainda é cedo para extrair conclusões, mas a julgar pelos resultados preliminares, a proposta tem grandes chances de se tornar um método seguro para a atribuição de descritores, com variados graus de verossimilhança representacional, dependendo de características como:

O campo de conhecimento de que tratam os textos;

A qualidade e atualização do tesauro utilizado;

O tamanho dos corpora analisados previamente;

Na medida em que mais resultados forem alcançados, serão divulgados nos fóruns apropriados.

6. Referencias

BAEZA-YATES, R.; **RIBEIRO-NETO**, B. *Modern Information Retrieval.* New York: ACM Press, 1999. 511p.

BICK, Eckhard. *Parsers and its applications*. (s/d) Disponível na Internet: http://www.hum.au.dk/lingvist/lineb/home_uk.htm. Consultado em 07/2003

_____. *Automatic parsing of Portuguese*. In: Proceedings of II Encontro para o Processamento Computacional do Português Escrito e Falado, SBIA, 1996, Curitiba. Disponível na Internet: http://beta.visl.sdu.dk/~eckhard/postscript/curitiba.ps. Consultado em 07/2003

_____. *The VISL System*: research and applicative aspects of IT-based learning. In: Proceedings of NoDaLiDa, 2001,Uppsala. Disponível na Internet: http://stp.ling.uu.se/nodalida01/pdf/bick.pdf. Consultado em 07/2003

GASPERIN, Caroline Varaschin; **GOULART**, Rodrigo Rafael Vilarreal e **VIEIRA**, Renata. *Uma Ferramenta para Resolução Automática de Correferência.* In: Anais do XXIII Congresso da Sociedade Brasileira de Computação, VI Encontro Nacional de Inteligência Artificial, Vol VII. Campinas, 2003.

GASPERIN, Caroline Varaschin; **VIEIRA**, Renata; **GOULART**, Rodrigo Rafael Vilarreal e **QUARESMA**, Paulo. *Extracting XML chunks from Portuguese corpora.* In: Proceedings of the Workshop on Traitement automatique des langues minoritaires. 2003. Batz-sur-Mer.

PROJETO DIRPI: Desenvolvimento e Integração de Recursos para Pesquisa de Informação. Cooperação Científica e Técnica Luso-Brasileira. ICCTI/GRICES-CAPES, Universidade de Évora, Universidade Nova de Lisboa, Unisinos, PUC-RS. Julho de 2001.

KURAMOTO, Hélio. *Uma abordagem alternativa para o tratamento e a recuperação de informação textual*: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, 1996. Disponível na Internet: http://www.ibict.br/cionline/250296/25029605.pdf. Consultado em 07/2003.

_____. *Proposition d'un Système de Recherche d'Information Assistée par Ordinateur Avec application à la langue portugaise*. 1999. Tese (Doutorado em Ciências da Informação e da Comunicação) – Université Lumière - Lyon 2, Paris, França.

LIBERATO, Yara G. *A Estrutura do Sintagma Nominal em Português:* uma abordagem Cognitiva. 1997. 203 f. Tese (Doutorado em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

MIORELLI, S. T. *Extração do Sintagma Nominal em sentenças em Português*. 2001. 98 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

MOREIRO, José; **MARZAL**, Miguel Ángel; BELTRÁN, Pilar. *Desarrollo de un Método para la Creación de Mapas Conceptuales*. Anais do ENANCIB, Belo Horizonte, 2003.

PERINI, Mário A. *A Gramática Gerativa*: introdução ao estudo da sintaxe portuguesa. 2ª edição. Belo Horizonte: Vigília, 1985. 254 p.

_____. *Gramática descritiva do português*. 2ª edição. São Paulo: Editora Ática, 1995. 380p.

PERINI, Mário A.; **FRAIHA**, Sigrid; **FULGÊNCIO**, Lúcia; **BESSA NETO**, Regina. *O SN em português*: A hipótese mórfica. **Revista de Estudos de Linguagem** - UFMG, Belo Horizonte, Julho / Dezembro 1996. p. 43-56.

ROSSI, Daniela; **PINHEIRO**, Clarissa; **FEIER**, Nara e **VIEIRA**, Renata. *Resolução automática de Correferência em textos da língua portuguesa*. REIC Revista de Iniciação Científica da SBC, http://www.sbc.org.br/reic/, v. 1, n. 2, 2001.

RUWET, Nicolas *Introdução à Gramática Gerativa*. São Paulo: Perspectiva, Editora da Universidade de São Paulo, 1975. 357 p.

SALTON, Gerard e **MCGILL**, Michael J. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company, 1983. 448 p.

SANT'ANNA, V. *Cálculo de referências anafóricas pronominais demonstrativas na língua portuguesa escrita*. 100 f. 2000. Dissertação (Mestrado em Informática) – Instituto de Informática da PUC-RS – Porto Alegre.

SPARCK JONES, K. e **WILLETT**, P. (orgs.). *Readings in Information Retrieval.* San Francisco: Morgan Kaufmann, 1997. 589p.

VIEIRA, R. *A review of the Linguistic literature on definite descriptions*. 1998. Acta Semiotica et Lingvistica, Vol. 7 : 219-258.

VIEIRA, R. et al. *Extração de Sintagmas Nominais para o Processamento de Co-referência*. 2000. Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada PROPOR, 19-22 Novembro Atibaia SP.

VISL. *About VISL*. Disponível na Internet: http://visl.hum.sdu.dk/visl/about/index.html. Consultado em 05/2003.