Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

EXPLICITAÇÃO DE ENTIDADES MENCIONADAS VISANDO O AUMENTO DA INTELIGIBILIDADE DE TEXTOS EM PORTUGUÊS

Marcelo Adriano Amancio
Sandra Maria Aluísio

NILC-TR-08-11
Agosto, 2008

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Muito se fala em inclusão digital, mas poucos projetos de tecnologia têm sido criados na área, em especial para o Português. Este Projeto de Graduação está inserido em um projeto maior financiado pela Microsoft Research-Fapesp, denominado PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), que almeja a construção de recursos e ferramentas que tornem textos mais inteligíveis. Como resultado, espera-se que analfabetos funcionais, usuários portadores de deficiências cognitivas ou pessoas em processo de alfabetização leiam textos da Web com menos dificuldades. Este projeto desenvolveu e avaliou intrinsecamente uma ferramenta computacional que visa o aumento da inteligibilidade em textos, um dos aspectos da acessibilidade. O aumento da inteligibilidade é conseguido através do reconhecimento automático de entidades mencionadas nos textos, para posterior explicitação visual e/ou explicação destas em um texto. Entidades Mencionadas (EM) são entidades referenciadas em um determinado contexto, por exemplo, são nomes próprios de pessoas, organizações, locais, acontecimentos, coisas (objetos nomeados), obras (artefatos e construções humanas), conceitos abstratos, além de datas e valores, podendo assim assumir papéis semânticos diferentes em função desse mesmo contexto. Usando características de três sistemas da literatura (Cortex Intelligence, Siemês, e Malinche), um protótipo foi desenvolvido e avaliado com as medidas estatísticas de precisão, abrangência (*recall*), e medida F que são tradicionais da área de recuperação de informação, área maior na qual se insere a tarefa de reconhecimento de EM. Para a sub-tarefa de identificação, a precisão conseguida foi de 66,64%, a abrangência de 72,16%. Para a sub-tarefa de classificação a precisão foi de 44.53% e a abrangência de 48.53%. Valores, estes, próximos dos resultados dos melhores sistemas participantes do HAREM.

1. Introdução

Este projeto se insere no escopo do projeto PorSimples (ALUÍSIO et. al., 2007). O pojeto PorSimples visa tornar textos da Web escritos em Português acessíveis para uma gama maior de usuários, usando métodos e técnicas de duas áreas de pesquisa: o Processamento de Língua Natural (PLN) e a área de pesquisa em Interação Usuário-Computador. Consideram-se, principalmente, os analfabetos funcionais, crianças em fase de alfabetização, e até mesmo pessoas com dificuldades cognitivas como os portadores de dislexia e afasia.

As técnicas estudadas no PorSimples são baseadas nos conceitos de Inteligibilidade e Acessibilidade. E, no contexto deste projeto de graduação I, será abordada uma técnica para aumentar a Inteligibilidade de textos, através da construção de uma ferramenta da área de PLN. Desta forma, serão estudados, revisados e avaliados métodos, técnicas e recursos do PLN, pois esta é a área de pesquisa na qual se inserem as pesquisas deste aluno.

Acessibilidade está diretamente relacionada com o desenvolvimento da sociedade de nossa época. Na era do computador fala-se no conceito de Acessibilidade Digital. Este conceito é responsável por dar condições de uso dos recursos disponíveis em meio digital a qualquer usuário interessado em acessá-los (TORRES, 2002). Quando nos referimos à Web, a acessibilidade corresponde à possibilidade de qualquer usuário, utilizando qualquer agente de software ou hardware que recupera e serializa conteúdo WEB, entender e interagir com o conteúdo de um *site*.

A Inteligibilidade está relacionada ao uso de palavras freqüentes e estruturas sintáticas menos complexas para facilitar a compreensão de um texto por um leitor, mas, atualmente, há uma preocupação com a avaliação da macroestrutura do texto além da microestrutura, em que outros fatores são vistos como facilitadores da compreensão, por exemplo, a organização dos textos, a coesão, a coerência, e o conceito do texto sensível ao leitor (LEFFA, 1996).

Há várias formas de se obter um aumento de inteligibilidade em textos. Três formas serão trabalhadas no escopo do PorSimples: (a) via explicitação das relações retóricas de um texto (por exemplo, identificando as relações de contraste, adição de idéias, causa-efeito, exemplificação, reformulações, etc.) (PARDO, 2005); (b) via explicitação de ações e sua estrutura argumental (isto é, estrutura de argumentos de alguns verbos e substantivos que necessitam desses argumentos, pois não têm auto-suficiência semântica (BORBA, 1996); (c) via explicitação de Entidades Mencionadas, que será o assunto deste projeto de graduação.

Entidades Mencionadas (EM) referem-se a nomes próprios encontrados em textos; também se consideram quantidades e referências temporais como fazendo parte de Entidades

Mencionadas (SANTOS, 2006b). O que se avalia em um sistema de REM é a identificação do sentido que as entidades representam no contexto em que estão inseridas e não todas as possibilidades de sentidos da entidade.

Neste projeto, aborda-se apenas a tarefa de Reconhecimento de Entidades Mencionadas (REM) e como resultado temos um protótipo operacional. A relevância desta pesquisa pode ser identificada pelas diversas aplicações, apresentadas na Seção 1.2.2, da tarefa de REM.

Em continuidade a este projeto seguirá o Projeto de Graduação II, que também abordará o aumento da inteligibilidade de textos escritos em Português. Porém, no Projeto II, será desenvolvido um identificador da estrutura argumental de verbos, com objetivo de aumentar a inteligibilidade das ações em textos escritos em Português. Detalhes do Projeto II podem ser encontrados no Apêndice A.

Há três formas de se avaliar uma tarefa de PLN: (a) usando-se medidas tradicionais como precisão e revocação/abrangência (*recall*), por exemplo, da tarefa realizada por um método automático com relação a uma anotação manual realizada em um córpus de referência (chama-se avaliação intrínseca); (b) por meio de uma avaliação de uma tarefa, sendo esta inserida em um sistema/aplicação maior para se avaliar o ganho da tarefa na aplicação; e (c) avaliação com usuários reais, para se avaliar também o ganho ou melhoria da tarefa via análise dela por usuários. A primeira forma é a mais usada, por ser mais rápida e menos custosa, pois não exige recurso humano de teste ou sistema/aplicação na avaliação do ganho.

1.1 Contextualização, Motivação e Domínio da Aplicação

1.1.1 A tarefa de REM

O termo Entidade Mencionada (NADEAU at. al., 2006) surgiu em 1996 na conferência *Message Understanding Conference* (MUC) (GRISHMAN et. al., 1996). Esta conferência, no seu papel principal de avaliação de sistemas de Recuperação de Informação, identifica a necessidade da tarefa de avaliação de sistemas REM.

A tarefa de REM pode ser resumida em identificar nomes próprios em textos e classificá-los em alguma classe semântica, pertencente a uma ontologia escolhida. Na tarefa REM comum, essa ontologia resume-se a: nome de organizações, pessoas, lugares, expressões numéricas, etc. Porém, é possível a construção de ontologias para domínios específicos, por exemplo, as Ontologias de localizações geográficas (MARTINS, 2006).

O processo de REM envolve, então, duas etapas principais. A primeira é a identificação de candidatos. Nesta fase, levantam-se todos os candidatos possíveis. A segunda etapa tem o papel de classificar as entidades e rever os passos realizados na etapa de identificação. Caso haja problema de ambigüidade podemos realizar a desambiguação em um passo adicional.

Dentre uma variada gama de aplicações de sistemas de REM, estes podem ser utilizados na explicitação da informação recuperada. Um exemplo funcional pode ser observado em um sistema da empresa Cortex Intelligence[1], cuja interface pode ser vista na Figura 1. Este sistema permite escolher textos para análise semântica (via botão "Sortear Novo"); explicitar as entidades mencionadas (via botão "Entidades"), explicitar as ações e sua estrutura argumental (via botão "Ações"). Podemos voltar ao texto original sem as explicitações de ações ou entidades via botão "Texto original".

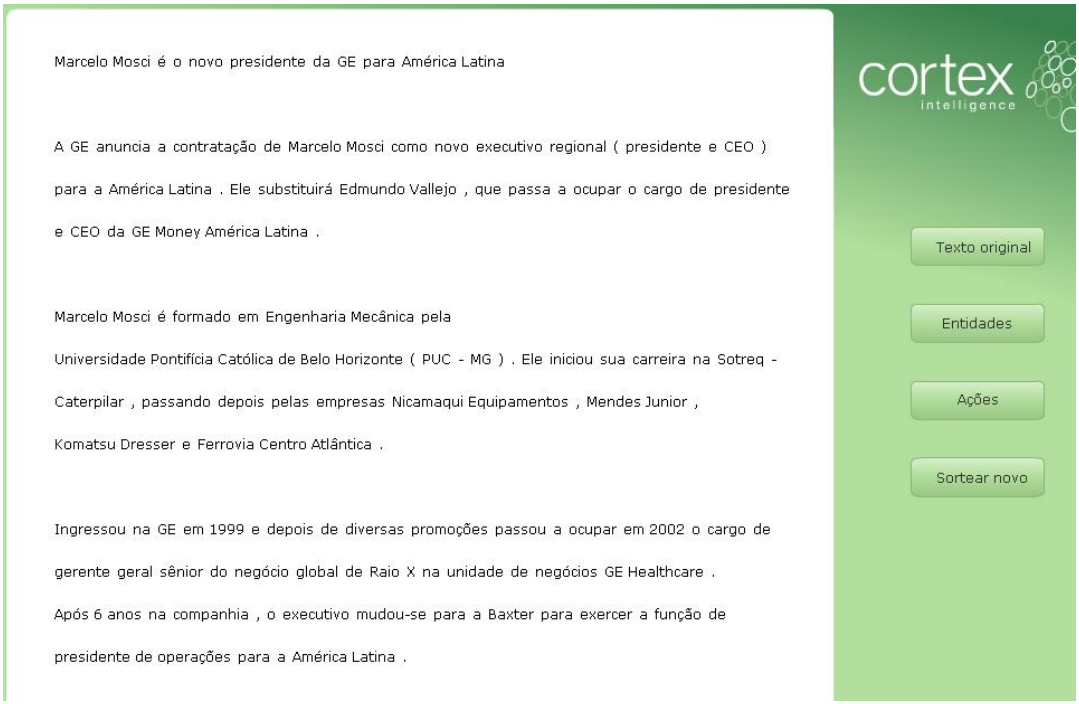

Figura 1: Interface principal do sistema Cortex

O trecho que será tomado como exemplo pode ser visto na Figura 2.

[1] http://www.cortex-intelligence.com/engine

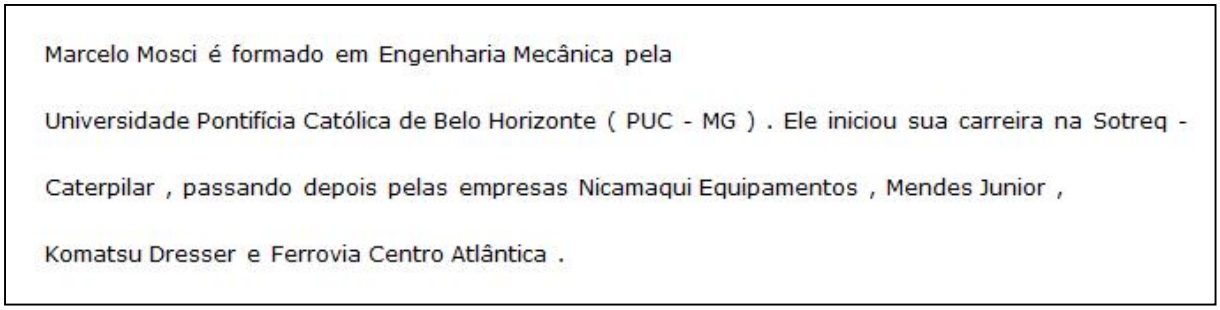

Figura 2: Texto anterior à anotação

O clique no botão *Entidades* destaca as Entidades Mencionadas, usando cores diferentes para cada classe semântica, como mostrado na Figura 3.

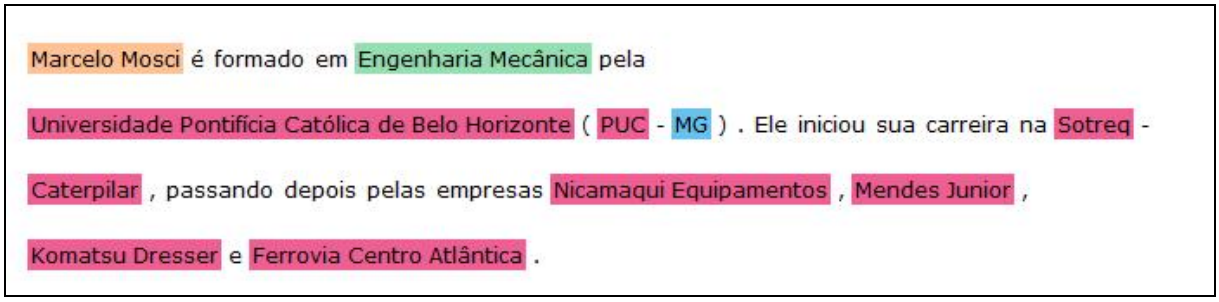

Figura 3: Texto com entidades mencionadas identificadas

Entidades similares, tais como *Nicamaqui Equipamentos* e *PUC*, nomes de instituições, foram explicitadas usando-se a mesma cor vermelha. Já o nome de pessoa, *Marcelo Mosci*, foi apresentado com a cor laranja. Ou seja, entidades semelhantes são, geralmente, diferenciadas por cores semelhantes.

Adicionalmente, há mais informações disponíveis ao usuário. Um clique sobre a sigla PUC (Figura 4) mostra uma nova janela com informações mais específicas sobre tal entidade (uma subcategorização e uma definição retirada da Wikipédia).

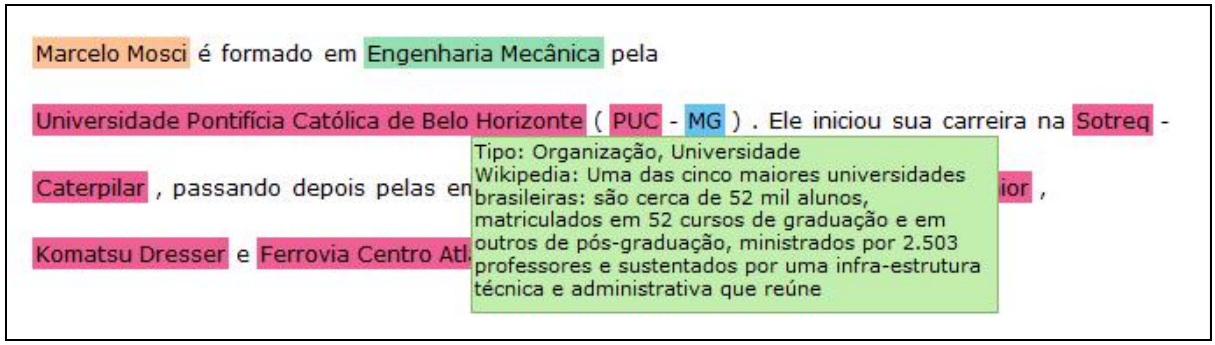

Figura 4: Informações detalhadas sobre a entidade PUC

Em suma, a idéia subjacente ao Reconhecimento de Entidades Mencionadas está em recuperar informação de uma maneira prática e rápida e explicitar ao usuário.

1.1.2 Aplicações

Os assuntos abordados tanto no Projeto de Graduação I como no II têm várias aplicações, em problemas atuais como antigos, mas que tem características em comum: a recuperação e enriquecimento da informação disponível, e grande potencial de impactar de forma positiva nas ferramentas atuais de recuperação da informação.

Além da aplicação no aumento da inteligibilidade em textos que é citada na Introdução, exitem várias outras. Na Seção 1.1.2.1, citamos a aplicação de aumento da velocidade de leitura que este projeto I, em conjunto com o projeto de graduação II, pode oferecer; a Seção 1.1.2.2 expõe-se aplicações na área de mineração de textos e inteligência competitiva, exploradas por poucas empresas brasileiras; na Seção 1.1.2.3 mostramos aplicações desta tecnologia nas áreas de Web Semântica e sistemas de Perguntas&Respostas; por fim, na Seção 1.1.2.4 apresentamos aplicações na área de tradução automática.

1.1.2.1 Leitura Veloz

Vários estudos têm sido utilizados em relação à construção de resumos (PARDO, 2002), (PARDO, 2003). A idéia por trás de um resumo é oferecer ao leitor um texto reduzido em relação ao original. Isto possibilita vários benefícios, como ler um texto de forma rápida e com menor necessidade de esforço cognitivo, uma vez que textos mais curtos, em geral, são mais acessíveis.

Acredita-se que os tópicos que serão abordados no projeto de graduação I e no II podem obter o mesmo resultado dos sumarizadores. Estudamos a possibilidade de comparar as duas abordagens no Projeto de Graduação II.

A representação visual de informação possibilita explorar a inteligência cognitiva visual humana, que é a proposta dos explicitadores. Um exemplo é dado no Apêndice A, no qual podemos observar que não é necessário ler uma frase toda para entender qual o assunto e os elementos mais importantes abordados. Notadamente, conseguimos identificar as principais ações em que os principais verbos ou substantivos estão envolvidos em apenas alguns segundos, apenas passando o olho sobre o texto, e identificando visualmente suas relações.

Dessa forma, o usuário pode direcionar seus esforços na observação das ações, relações e nomes, pois são estas as entidades que proporcionam o entendimento. E, se necessário, o leitor pode ler o texto em sua integridade. Ou de forma mais inteligente, só a parte mais importante.

1.1.2.2 Mineração de textos e Inteligência Competitiva

Na tese de ARANHA (2007a), mostra-se a necessidade do Reconhecimento de Entidades no pré-processamento da mineração de textos. Tal processo advém da necessidade de dar nomes aos objetos do mundo real em questão.

Também é mencionado em ARANHA (2007a) que a maioria das informações de uma notícia de jornal, por exemplo, provém principalmente da combinação de dois ou mais nomes conhecidos. Calcula-se que essa freqüência se dê em 90% desses novos lexemas[2].

1.1.2.3 Web Semântica (Web 3.0) e sistemas de Perguntas&Respostas

A identificação de classes de nomes e relacionamentos de palavras é uma tarefa comum ao ser humano, mas não natural a um computador. A Web de hoje e seus sistemas de busca ainda não usufruem da capacidade que os humanos têm de realizar estas tarefas cognitivas.

Quando necessitamos de uma informação, geralmente usamos estes buscadores. As buscas disponibilizadas atualmente funcionam através da requisição de uma seqüência de palavras (termos) junto a um buscador e obtenção da resposta, uma lista de web sites ordenados. Esta forma de busca tradicional não identifica as intenções dos usuários, apenas retornam um conjunto de sites relacionados com o termo procurado.

A Web Semântica promete superar as limitações atuais e será provavelmente a nova geração dos buscadores.

[2] Um lexema corresponde às formas (morfologia) que uma palavra se dispõe.

O Reconhecimento de Entidades Mencionadas, a identificação da estrutura argumental de verbos e outros recursos lingüísticos similares fazem parte do processo de obtenção de uma Web com informações semânticas. Com informações semânticas temos uma maior capacidade de poder aplicar uma busca mais voltada às intenções dos usuários. Um sistema nesta direção pode ser visto em PEREIRA (2007). Um processo similar é usado em sistemas de Perguntas&Respostas (BABYCH et. al., 2003).

1.1.2.4 Tradução de Máquina

A aplicação de sistemas de REM em textos fornece um enriquecimento semântico de seu conteúdo. Tradutores necessitam de informações semânticas dos elementos do texto para diminuir sua possibilidade de erro.

Uma nova tendência em tradução de termos é vista no comportamento dos usuários. Esta tendência é observada na tradução de nomes utilizando o auxílio da enciclopédia on-line Wikipédia. Suponha que procuramos o termo em inglês *Artificial Intelligence*. Dada esta busca poderemos obter mais de 50 artigos nas mais diversas línguas, cujo título será o termo traduzido na língua destino. Mais detalhes deste processo pode ser obtido em MIHALCEA et. al. (2001).

1.2 Objetivos do Trabalho

No contexto do Projeto de Graduação I, realizou-se um estudo de sistemas de Reconhecimento de Entidades Mencionadas. Dos principais sistemas conhecidos atualmente para a língua portuguesa estão aqueles participantes do campeonato internacional de Reconhecimento de Entidades Mencionadas, o HAREM (CARDOSO et. al., 2007d). A escolha dos sistemas abordados teve como princípio observar a diversidade de técnicas que podem ser usadas na tarefa, como também dar embasamento à construção de um protótipo de sistema REM. Também, foi levantada a arquitetura básica da maioria dos sistemas encontrados na literatura. Métricas de avaliação de sistemas REM são apresentadas. Dentre elas, as utilizadas na competição HAREM. Dadas as métricas e os cenários de avaliação, o sistema desenvolvido foi avaliado.

1.3 Organização da Monografia

Na Seção 2 é feita uma revisão bibliográfica, não apenas dos sistemas de REM existentes, mas também de recursos, e de um campeonato de avaliação de sistemas da área. Quatro sistemas são detalhados na Seção 2.3 e suas avaliações comparadas na Seção 2.3.5.

Na Seção 3 fornecemos uma exposição mais detalhada do protótipo de um sistema de REM desenvolvido, apresentando sua arquitetura e avaliações. As experiências adquiridas no estudo de sistemas REM e na construção do protótipo dão embasamento à discussão das limitações da tarefa e sugestões de melhoria, que podem ser vistas em Seção 3.3.6 e Seção 3.4.

Na Seção 4 é apresentada a conclusão do projeto. Na Seção 5 são dadas algumas sugestões para o curso de graduação.

2. Revisão Bibliográfica

Na Seção 2.1 são dados os conceitos, terminologias e alguns recursos usados em sistemas de REM. Na Seção 2.2 apresenta-se uma visão sobre como estes conceitos e recursos se relacionam. Na Seção 2.3 quatro sistemas REM são descritos e são comparados em relação as suas pontuações no campeonato HAREM.

2.1 Conceitualização e Terminologia

Entidades Mencionadas (EM) são palavras da classe de substantivos próprios. Definem nomes para lugares, pessoas, organizações, etc. Reconhecimento de Entidades Mencionadas (REM) é a tarefa de localizar e explicitar as Entidades Mencionadas em um texto. Essa explicitação significa o enquadramento dessa entidade em uma ontologia específica, previamente estabelecida.

Córpus é um conjunto de textos que são compilados para um fim específico. É um recurso lingüístico na atividade de REM.

Gazetteers/Almanaques são dicionários de Entidades Mencionadas. Podem ser compilados de forma semi-automática, automática ou manual. São geralmente construídos para um fim específico (NADEAU, 2006).

Metapalavras, nomenclatura herdada de ARANHA (2007b), representam as palavras das vizinhanças das entidades. Estas palavras muitas vezes dão indicações das classes destas entidades. Por isso, Metapalavras são geralmente usadas na etapa de desambiguação. Exemplos: rio, jogador, avenida.

Adivinhação, nomenclatura herdada de ARANHA (2007b), são similares às Metapalavras. Adivinhações são palavras que fazem parte das entidades e também dão dicas de sua classificação. São usados na etapa de classificação. Exemplo: Dr., Sr.

Regras de similaridade são um conjunto de regras, geralmente codificados em expressões regulares, que definem similaridades entre a entidade a ser classificada e as entidades existentes no Gazetteer/Almanaque. Regras de similaridade e almanaques têm o mesmo resultado de um córpus anotado com EM, em conjunto com aprendizado de máquina.

2.1.1 Repentino

Repentino[3] é um gazetteer, isto é, um dicionário de Entidades Mencionadas que foi compilado no pólo de pesquisas Linguateca (SARMENTO et. al., 2006b). Este grupo de pesquisadores, na construção do sistema Siemês, identificou a escassez deste recurso lingüístico (gazetteer) para a língua Portuguesa e iniciou a tarefa de construir seu próprio Gazetteer, que resultou no Repentino. Atualmente, o gazetteer Repentino é disponível publicamente[4]. Este gazetteer é composto por mais de 450 mil entradas. Optou-se pela generalidade de tópicos em relação a aprofundar-se em poucos deles. Com isso, Repentino é dito ser um gazetteer de escopo amplo. Apesar de ser compilado de forma semi-automática, teve uma validação totalmente manual. Isto torna este recurso valioso por sua precisão humana neste processo. Podemos ver sua ontologia no Anexo A. O gazetteer Repentino é usado no protótipo deste projeto de graduação.

2.1.2 Coleção Dourada

A Coleção Dourada (CD) é um córpus criado para o uso da avaliação dos sistemas no HAREM (Seção 2.1.3). Foi concebido um córpus de textos de vários gêneros e, neste córpus, vários anotadores humanos se incumbiram de anotar os textos seguindo as diretivas de Reconhecimento de Entidades Mencionadas disponibilizadas. Recursos com as características da CD são importantes tanto na avaliação de sistemas da área, quanto em seu uso ou como recurso lingüístico no aprendizado de sistemas REM.

[3] Http://www.linguateca.pt/repentino/
[4] http://poloclup.linguateca.pt/repentino/repentino.xml.gz

2.1.3 HAREM

HAREM (Avaliação de Reconhecimentos de Entidades Mencionadas) é o primeiro campeonato de avaliação de Sistemas de REM para o Português. É realizado pelo pólo de pesquisas Linguateca (Oliveira et. al., 2003). Os objetivos da criação e continuidade deste campeonato são:

1 – Ajudar a comunidade científica em concordar nos requisitos mínimos da tarefa proposta;

2 – Conhecer a comunidade atuante na área;

3 – Avaliar sistemas REM, e eventuais opções na sua implementação, de forma independente de estrutura;

4 – Obter recursos valiosos para avaliação no futuro, tal como a Coleção Dourada.

A metodologia de avaliação baseia-se na comparação do resultado dos sistemas participantes em relação a um córpus anotado por humanos – a Coleção Dourada (CD). A validação estatística deste formato de avaliação pode ser encontrada em CARDOSO et. al. (2007b).

Para que houvesse uma padronização na anotação destes sistemas, foi criado um conjunto de diretivas, que podem ser encontradas em CARDOSO et. al. (2007a). O conteúdo deste documento refere-se desde aos formatos das etiquetas a serem usadas até quais entidades devem ser identificadas e quais características devem ser consideradas em sua classificação

A avaliação realizada no HAREM é denominada avaliação conjunta (CARDOSO et. al., 2007c); este termo será usado neste texto. Os sistemas inscritos podem participar em categorias específicas. Por exemplo, se temos um sistema que só reconhece entidades mencionadas geográficas, podemos optar por termos uma avaliação restrita a este cenário. A nomenclatura usada em CARDOSO et. al. (2007c) é a de avaliação em cenários seletivos.

2.2 Recursos usados no Reconhecimento de Entidades Mencionadas

A arquitetura de sistemas de REM é geralmente composta por um conjunto de regras de comparação e análise de Entidades Mencionadas. Estas regras atuam sobre um conjunto de recursos lingüísticos, tal como gazetteers, metapalavras, etc. Este processo tem como objetivo analisar textos, usando-se recursos relevantes no processo de anotação. Nesta análise são exploradas tanto dicas internas das entidades, tal como dados ortográficos, por exemplo,

palavras iniciadas por maiúsculas/minúsculas, todas as letras são maiúsculas (siglas), quanto dicas externas, exemplos de metapalavras.

Gazetteers ajudam na tarefa de identificação e classificação correta das entidades uma vez que este recurso possui um grande conjunto de exemplos corretos. Nesses exemplos podemos encontrar desde a entidade, quanto o contexto em que ela foi encontrada. Isto possibilita a aplicação de regras internas quanto externas, explicadas anteriormente.

Córpus anotados também podem ser usados nas tarefas de REM, pois contêm todos os recursos lingüísticos necessários a este fim. Junto a um córpus, geralmente são usadas estratégias de aprendizado de máquina. Uma das principais dificuldades que desfavorecem o uso de córpus é a sua difícil obtenção. Necessita-se de um grande esforço humano para a sua construção, apesar de ser um recurso que fornece bons resultados. Gazetteers tem vantagem de serem construídos de forma semi-automática, uma vez que padrões de busca (expressões regulares) podem ser usados, reduzindo a dificuldade da construção de tal recurso. Geralmente adota-se a estratégia de aprendizado de máquina em um córpus anotado, ou então, um gazetteer e um conjunto de regras. Podemos ver o uso do aprendizado de máquina na Seção 2.2.3, e de uso de regras na Seção 2.2.2.

Na etapa da desambiguação, a técnica mais adotada é o uso de regras. Estas regras, em geral, não são muito extensas e são compiladas manualmente. O sistema Siemês usa esse conjunto de regas.

2.3 Trabalhos Relacionados

Nesta seção, quatro sistemas de REM são abordados. Na Seção 2.2.1 é descrito o sistema Cortex, na Seção 2.2.2 descrevemos o sistema Siemês, na Seção 2.2.3 o sistema Malinche e na Seção 2.2.4 o sistema CaGe. Por fim, em Seção 2.2.5 são apresentadas as pontuações dos sistemas no campeonato HAREM.

2.3.1 Sistema Cortex Intelligence

O sistema Cortex Intelligence foi originalmente desenvolvido por Christian Nunes Aranha durante o desenvolvimento de sua tese de Doutorado (ARANHA, 2007a) na PUC-Rio. Esta primeira versão participou do campeonato HAREM (ARANHA, 2007b) e obteve o primeiro lugar geral (dados não oficias) (ARANHA, 2007b). Atualmente, este sistema é mantido pela empresa *Cortex Intelligence*[5], sua última versão é a 3.0.

[5] http://www.cortex-intelligence.com/engine

O molde teórico adotado baseia-se na modelagem das faculdades cognitivas humanas. Adaptabilidade, flexibilidade, antecipação, pressuposição e revisão de hipóteses são refletidas em cada passo dos processos do sistema. É o mesmo mecanismo que uma criança, na espontaneidade de seu aprendizado, adquire a fala.

O resultado desta modelagem é a capacidade do aprendizado com novos textos, e dessa forma, diminui sua dependência de recursos estáticos[6], tal como gazetteers ou enciclopédias. Já um ponto negativo é a sua dependência de língua (atualmente, o sistema só trabalha com o Português), porém, de forma similar a um ser humano, o sistema é capaz de aprender novas línguas.

A arquitetura do sistema (Figura 5) adota o uso de dicionários, enciclopédias, a modelagem da gramática da língua.

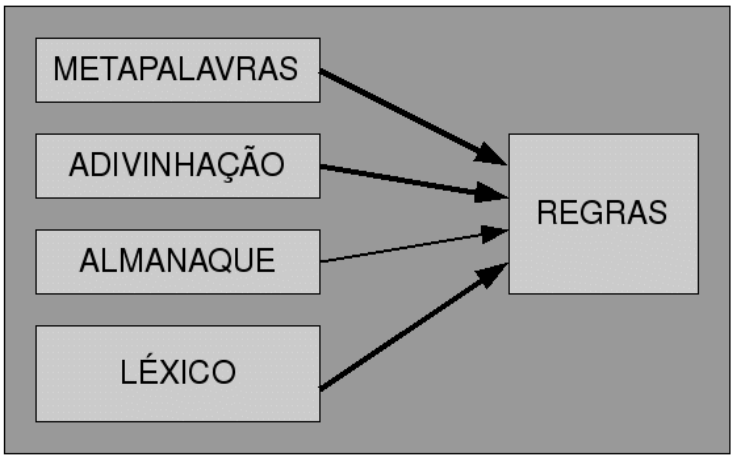

Figura 5: Arquitetura de items de auxílio no Cortex

A arquitetura mostrada resume-se à utilização de quatro recursos (metapalavras, adivinhação, almanaque e léxico) na criação de suas regras de REM. Em suma, o sistema Cortex não apenas realiza uma boa identificação na tarefa REM como também aprende com novos textos que lê (PEREIRA, 2007).

2.3.2 Siemês

Siemês (SARMENTO, 2006a) é um sistema de REM desenvolvido pela Faculdade de Engenharia da Universidade de Porto (NIAD&R) e Linguateca. Este sistema é baseado em um gazetteer, o Repentino (SARMENTO, 2006a), e em cinco regras de similaridade. Esse

[6] Este comportamento é dito criativo, pois usa o conhecimento aprendido em relação a dados estáticos, aumentando a sua base de conhecimento.

conjunto permite a obtenção de hipóteses de classificação usando-se apenas evidências internas, que poderão ser desambigüizadas posteriormente, a partir de um conjunto de regras simples, baseadas em dicas de contexto. Segue a arquitetura de três estágios na Figura 6.

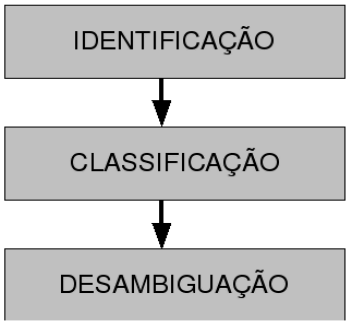

Figura 6: Arquitetura do Siemês

Na identificação são usadas informações ortográficas (maiúscula), ou a presença de números. A presença de uma dessas duas informações é considerada como sementes, que nos próximos passos podem adicionar palavras da vizinhança em sua entidade. Suas regras almejam encontrar similaridades entre a suposta entidade identificada e a entidade existente no Repentino. As cinco regras são as seguintes, que são:

1 – As duas entidades não diferem em nenhum caractere (entidades iguais);

2 – As duas entidades iniciam da mesma forma (substrings iniciais iguais);

3 – As duas entidades terminam da mesma forma;

4 – Se entidades possuem similaridades internas;

5 – Verificação se a entidade possui uma palavra freqüente de uma dada subclasse.

A desambiguação, terceira etapa do Siemês, é efetuada usando-se regras manualmente compiladas. Tais regras podem desambiguizar dois candidatos cujas ambigüidades foram levantadas no estágio anterior e serão resolvidas nesta etapa, ou também, podem achar um candidato que passou não classificado pela etapa anterior.

2.3.3 Sistema Malinche

Malinche é um sistema REM que foi originalmente construído para identificação de entidades do espanhol. Atualmente, ele adapta seus métodos para alcançar novas línguas (SOLORIO, 2005), inclusive o Português.

Este sistema adota o uso de um córpus[7] em conjunto com técnicas de aprendizado de máquina. Usa a arquitetura de dois estágios, a seguir:

A delimitação de entidades mencionadas que associa à cada palavra um identificador que diferencie essas entidades. Os identificares possíveis são:

B – identificador para palavras que iniciam uma entidade;

I – associa palavras que pertencem a uma entidade, mas não são seu início;

O – É a exclusão das possibilidades acima, ou seja, aquelas palavras que não são nem B nem I.

Essa associação é realizada utilizando um conjunto de treino previamente anotado. A Tabela 1 ilustra como é feita a adaptação do córpus (pré-anotação) no esquema de etiquetas mais simples (BIO). É este novo conjunto a entrada para o algoritmo de delimitação.

Tabela 1: Sentença anotada no formato BIO

Palavras	**Pré-anotação**	**Anotação BIO**
Barack	Nome	B
Obama	Nome	I
candidata-se	-	O
à	-	O
presidência	-	O
dos	-	O
EUA	País	B

Observa-se que este conjunto é independente do algoritmo de aprendizado de máquina. Porém, o sistema Malinche usa o algoritmo Support Vector Machines (Vapnik, 1995). Este algoritmo é especialmente bom para dimensionalidades altas, pois permite a divisão do espaço de maneira não linear, em um tempo computacional satisfatório. Na tarefa de classificação o sistema usa cinco atributos para cada palavra:

1- Informação ortográfica: se a palavra tem minúsculas, maiúsculas ou envolve dígitos;

[7] O córpus utilizado foi desenvolvido pelo grupo de pesquisas TALP (Carreras et al., 2002) e foi anotado à mão.

2 - A posição da palavra na sentença;

3 - A própria palavra;

4 - A anotação completa;

5 – E a anotação reduzida (tag BIO).

O algoritmo de treinamento usa estes atributos da palavra atual em seu processo de treinamento. Ao contrário de outros sistemas, o processo adotado não usa recursos lingüísticos tais como listas de palavras, dicionários, gazetteers e expressões de contexto. O uso de um córpus anotado é rico o bastante para substituir recursos lingüísticos extras, se este último for grande o suficiente.

Este sistema demonstra resultados bons e também portabilidade para novas línguas. O ponto crítico do sistema é o córpus, uma vez que é muito custoso manter um de alta qualidade, pois exige recurso humano caro. Em contrapartida, permite que os métodos adotados sejam relativamente independentes de língua.

2.3.4 Sistema Cage (Capturing Geographic Entities)

O sistema CaGE (MARTINS et. Al, 2007) é um sistema REM de Entidades Geográficas. O Reconhecimento de entidades geográficas faz parte de muitos sistemas SIGs[8], que também trabalham com dados não estruturados, tal com em textos.

O reconhecimento de Entidades Geográficas tem algumas peculiaridades e por isso é abordado nesta monografia. Um dos problemas encontrados nesta tarefa é a identificação unívoca de entidades. Por exemplo, um nome pode corresponder a mais de um lugar, ou um lugar pode ter mais de um nome.

A ambigüidade pode ser resolvida no armazenamento da informação na criação de uma ontologia completa não ambígua, que considere, por exemplo, nome de cidade, estado, país, etc. Já na fase de identificação, o problema persiste, exigindo elementos de contexto.

O sistema CaGE usa duas ontologias geográficas, uma é baseada em informações de nomes geográficos em nível de mundo, e outra usa em nomes de dados geográficos específicos de Portugal[9]**.** Tais entidades usam nomes de locais, relações topológicas, dados geográficos. Junto a este recurso são usadas regras de contexto correspondente à existência de

[8] Sistemas de Informações Geográficas

[9] O sistema CaGE é usado no website www.tumba.pt site de buscas portugues. Por isso a consideração de informações geográficas portuguesas.

maiúsculas, expressões, nomes de locais e tabela de exceções. Cada qual visa à identificação eficiente e conseguinte desambiguação das entidades geográficas presentes no texto.

CaGE é um sistema de 4 estágios: pré-processamento, identificação, desambiguação e geração de anotações.

O pré-processamento corresponde à conversão de uma página Web em texto (não utilizado no contexto do HAREM), também a atomização de palavras e reconhecimento de frase baseado em "pares de contexto", e por fim, à divisão das frases em seus n-gramas correspondentes.

A etapa de identificação localiza n-gramas que possivelmente correspondem a entidades geográficas. Essa identificação é realizada através da comparação de informações de letras maiúsculas, expressões, e ontologia consideradas.

Por fim, a desambiguação é realizada na última etapa e corresponde à aplicação de regras de classificação, classificação baseadas em ontologia, comparação de referências ambíguas e ordenação de conceitos geográficos.

A geração de anotações é a parte que anota a saída com formatações previamente especificadas, resultando em formato similar XML.

2.3.5 Comparando os sistemas

Nesta seção, são comparados os resultados dos sistemas previamente descritos no HAREM 1. Estas informações foram obtidas do texto (SANTOS et. al., 2006b). A Tabela 2 mostra os resultados destes sistemas na tarefa de identificação, com os melhores resultados sendo do Siemês.

Tabela 2: Resultados da tarefa de Identificação

	Precisão (%)	Abrangência (%)	Medida F
Cortex	62,68	43,51	0,5136
Siemês	**77,15**	**84,35**	**0,8059**
Malinche[10]	-	-	-
CaGE	51,61	12,28	0,1984

[10] Sistema não participante

Na Tabela 3 são apresentados os resultados da tarefa de classificação semântica, com os melhores resultados sendo do Siemês.

Tabela 3: Resultados da tarefa de Classificação

	Precisão (%)	Abrangência (%)	Medida F
Cortex	47,95	33,09	0,3916
Siemês	**57,28**	**49,85**	**0,5630**
Malinche	-	-	-
CaGE	36,22	6,85	0,1152

Como podemos notar, é quase sempre mais difícil conseguir abrangência do que precisão. E, um bom equilíbrio em ambas possibilita uma boa pontuação na medida F, como foi o caso do sistema Siemês na tarefa de identificação. O sistema Córtex apresenta maior pontuação atualmente. Seu resultado oficial é baixo, pois teve penalizações na pontuação de seu sistema por problemas de adaptação do formato da saída de seu programa.

3. Estado atual do trabalho

Nesta seção, os tópicos relacionados ao desenvolvimento deste projeto são abordados. Na Seção 3.1 são definidas as arquiteturas do protótipo em questão; na Seção 3.2 apresenta-se a equipe do projeto maior – o PorSimples; na Seção 3.3 são descritas as tarefas realizadas; na Seção 3.4 mostram-se os resultados obtidos; e, por fim, na Seção 3.5, é feito um levantamento das dificuldades encontradas.

3.1 O projeto

O projeto desenvolvido envolve a construção de um protótipo de Reconhecimento de Entidades Mencionadas. Este protótipo é validado usando a Coleção Dourada[11] do HAREM I.

Este protótipo é construído junto a outros recursos que possibilitam a sua avaliação. Os módulos do projeto podem ser vistos na Figura 7.

[11] http://poloxldb.linguateca.pt/HAREM.php?l=coleccaodourada

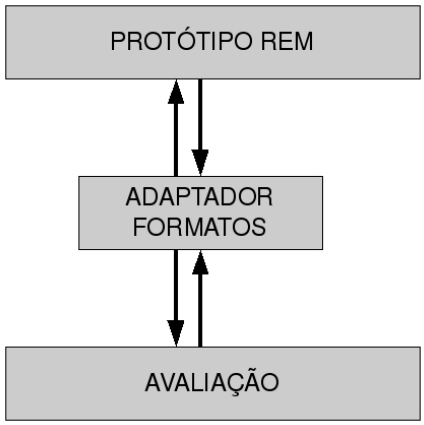

Figura 7: Módulos do Sistema

O primeiro módulo - o Protótipo REM - realiza o processamento da tarefa REM em textos. O módulo Avaliação levanta as medidas de abrangência, precisão e medida-F do sistema e avalia o sistema. Já o Adaptador de Formatos realiza o intercâmbio de formatos.

O primeiro passo do Adaptador de Formatos é duplicar a coleção dourada. Após a duplicação, uma das cópias é mantida intacta, e a outra tem suas etiquetas removidas e é submetida ao protótipo. O Protótipo anota este conjunto de textos e retorna-os ao módulo Adaptador de Formatos. O Adaptador de Formatos adiciona novas etiquetas a este córpus, para que ele seja comparado com o conjunto de textos mantidos intactos.

Com os dois conjuntos de textos obtidos na etapa anterior podemos enviá-los ao módulo de Avaliação e obter as médias de precisão, abrangência e medida-F. A arquitetura do

Com os dois conjuntos de textos obtidos na etapa anterior podemos enviá-los ao módulo de Avaliação e obter as médias de precisão, abrangência e medida-F. A arquitetura do protótipo pode ser vista na Figura 8.

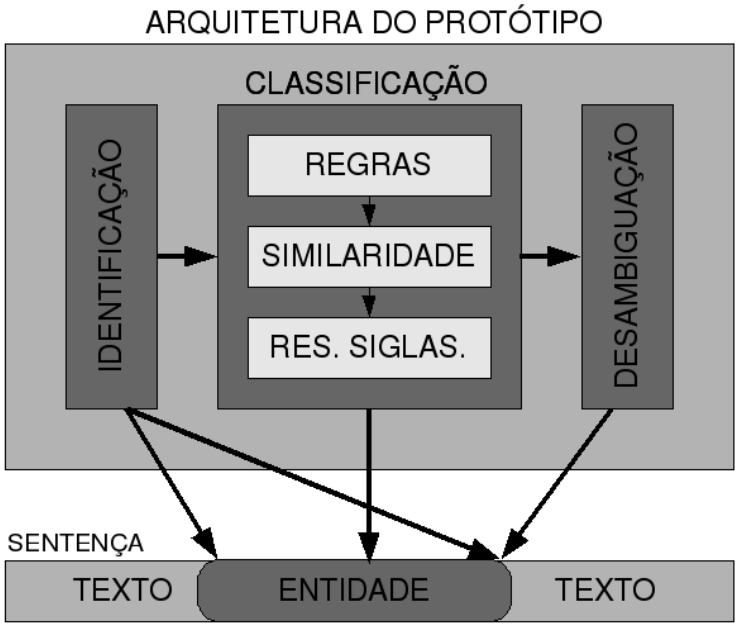

Figura 8: Arquitetura do protótipo desenvolvido

A Figura 8 representa uma arquitetura de 3 camadas: 1 – identificação; 2 – classificação; e 3 – desambiguação. Estas camadas serão detalhadas nas próximas seções.

3.1.1 Identificação

Na etapa de identificação usamos dicas ortográficas como principal fonte de informação. Siglas são anotadas usando-se expressões regulares.

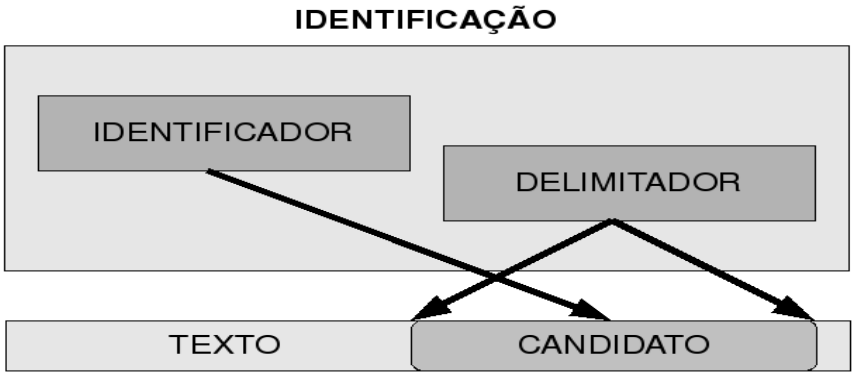

Figura 9: Módulos da etapa de Identificação

A Figura 9 mostra os módulos da etapa de Identificação. O primeiro módulo, o Identificador, consiste em identificar a presença das entidades no texto. Já o próximo módulo, o Delimitador, encontra as fronteiras do candidato.

3.1.2 Classificação

A classificação pode ser dividida em três etapas: regras (expressões regulares); similaridades; e resolução de siglas.

Expressões Regulares são usadas para a identificação de padrões que geralmente não tem muita variação. Entre eles podemos enumerar seis tipos de entidades:

1 - Números ou quantidades (números escritos por extenso, números inteiros, reais, etc.);

2 – Medidas (Anos, Meses, Dias, Horas, Minutos, Segundos, Medidas Numéricas, Medidas de Distância como Km, metros, etc.);

3 – Classificação (º, ª, primeiro, primeira, etc.);

4 – Medida Monetária (Real, Euro, etc.);

5 – Datas;

6 – Localizações virtuais (e-mail, por exemplo).

Após efetuar a aplicação de regras simples, necessitamos de regras mais gerais. Para isso usamos o gazetteer Repentino. Com ele, vamos utilizar três regras de comparação entre a estrutura identificada e suas entradas.

O primeiro passo é o processo de normalização. A normalização é um processo muito utilizado na área de Recuperação de Informação. Esta tarefa resume-se em eliminar detalhes não importantes das palavras, no processo em questão. A normalização utilizada neste caso corresponde à:

1 – Eliminação das diferenças ortográficas. Ou seja, transformação de todos os caracteres em minúsculas.

2 – Eliminação de espaços brancos em excesso.

3 – Eliminação de acentos, visto que há a possibilidade de encontrarmos palavras com acentuação errônea.

4 – Eliminação de cedilha, visto que em alguns textos os usuários podem substituir o cedilha pela letra “c”, ou vice-versa.

5 – Eliminação de caracteres não alfa-numéricos. Tais como: dois pontos, vírgulas, ponto e vírgula, etc.

Poderia ter-se optado por utilização do radical das palavras, mas notou-se que muita informação poderia ser perdida.

Esta normalização é efetuada tanto na entrada do gazetteer quanto nas entidades a serem classificadas. Logo a seguir, as entidades são classificadas usando-se três regras:

1 – A primeira regra é igualdade completa entre as entidades. Se há correspondência caractere a caractere entre a entidade a ser classificada e a entrada do gazetteer então a primeira entidade é classificada com a classe da entrada correspondente.

2 – A regra dois só ocorre se a regra 1 falhar. Esta etapa corresponde a verificar inícios de frase. Se a entidade do gazetteer e a entidade a ser classificada iniciarem da mesma forma então esta entidade é classificada com a classe da entrada do gazetter.

3 – A regra três é similar à regra dois, com a diferença que a similaridade encontra-se no fim das entidades.

A última etapa do processo de classificação é a resolução de siglas. Esta etapa equivale a fazer um levantamento das potenciais siglas do texto atual. Por exemplo, no texto "O período de inscrição ao Enem 2008 (Exame Nacional do Ensino Médio) foi estendido até 13 de junho", temos a sigla "Enem" e a sub-frase "Exame Nacional do Ensino Médio", que pode ajudar a obter a classificação correta da sigla.

3.1.2 Desambiguação

A última etapa é a desambiguação. Para ela, foi construída uma estrutura que possibilitasse a definição da ambigüidade e sua possível desambiguação.

A primeira delas corresponde à definição de regras mais gerais. O formato da regra é especificado por três valores: <DISTANCIA> <PALAVRA> <CLASSE>.

<DISTANCIA>: Corresponde ao número de palavras que <PALAVRA> precede <CLASSE>

<PALAVRA>: É a palavra da vizinhança da entidade que precede a entidade.

<CLASSE>: É a nova classe.

Exemplo: rua de <nome tipo="santo">santa luzia</nome>
Regra de desambiguação: 2 rua LOCAL
Resultado: rua de <local tipo="fixo">santa luzia</local>

3.2 Equipe

Este trabalho faz parte do projeto "PorSimples: Simplificação Textual do Português para Inclusão e Acessibilidade Digital", aprovado no âmbito do Edital Microsoft-Fapesp (proc. nro. 2007/54565-8) (Aluísio et. al., 2007), formado por pesquisadores do NILC e do laboratório Intermídia, ambos do ICMC. Ao todo a equipe conta com 19 integrantes. Mais detalhes da equipe (ou também na Wiki[12]):

(i) *Seis pesquisadores seniores*: Sandra Maria Aluísio (coordenadora do projeto – PLN), Lucia Specia (colaboradora – PLN), Maria da Graça Pimentel (colaboradora- HCI), Maria das Graças Volpe Nunes (colaboradora – PLN), Renata Fortes (colaboradora – HCI) e Thiago Pardo (colaborador – PLN).

(ii) *Uma pós-doutoranda*: Helena de Medeiros Caseli (PLN).

(iii) *Cinco pós-graduandos*: Amanda Rocha Chaves (PLN), Arnaldo Candido Jr. (PLN), Fernando Muniz (PLN), Marcelo Adriano Amancio (PLN), Willian Watanabe (HCI).

(iv) *Sete alunos de iniciação científica*: Carolina Scarton (PLN), Erick Galani Maziero (PLN), Felipe Vianna Perez (HCI), Gabriel Muniz Antonio (HCI), Henrique Valim Gnann (HCI), Paulo Rodrigues Alves Margarido (PLN), Tiago de Freitas Pereira (PLN).

3.3 Descrição das atividades realizadas

A seguir, são descritas as atividades realizadas neste projeto. Na Seção 3.3.1 é apresentada a arquitetura básica de um sistema de REM. Na Seção 3.3.2 são descritas as medidas utilizadas. Na Seção 3.3.3 é feita a avaliação do protótipo. Por fim, na Seção 3.3.4 são levantadas dificuldades e dadas sugestões de melhoria.

3.3.1 Levantamento de uma arquitetura básica

Tendo em vista o estudo dos sistemas abordados na Seção 2.3, podemos levantar a arquitetura comum mais utilizada por sistemas de REM. Aqui é descrita a arquitetura básica, além de algumas alternativas de implementação encontradas. Sabe-se que o objetivo da tarefa de Sistemas de Reconhecimento de Entidades Mencionadas é identificar, delimitar e classificar entidades. Portanto, todo sistema REM obedece uma arquitetura de no mínimo dois passos, mostrada na Figura 10.

[12] http://caravelas.icmc.usp.br/wiki/index.php/Principal

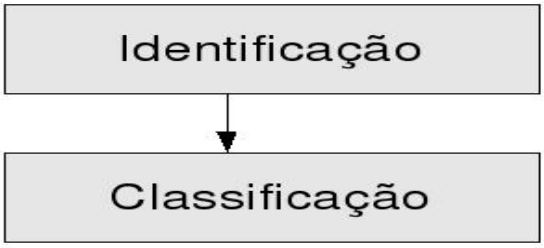

Figura 10: Processo Básico REM

No primeiro passo realiza-se a identificação de candidatos de Entidades Mencionadas, a seguir, segue-se o processo de identificação de classes nessas supostas entidades.

Como veremos mais adiante, os processos de identificação e classificação têm suas similaridades, uma vez que na etapa de classificação os passos da identificação podem ser revistos.

No processo de identificação é permitido identificar entidades que não sejam corretas, ou identificar as corretas com seus limites errados, uma vez que erros de excessos de anotações serão revistos mais adiante. Apesar de permitido, deseja-se que esses erros sejam minimizados. Em contraste, uma ausência de identificação em um candidato em potencial pode ser muito indesejável, pois o classificador não será capaz de rever tal processo, e reverter este tipo de erro.

As técnicas aqui usadas, na maioria das vezes, recorrem a dicas de ortografia, basicamente verificando se um nome inicia-se com letras maiúsculas. Porém este recurso não cobre todos os casos, exemplo disto são inícios de frase.

Em suma, a Identificação pode ser composta nas tarefas de Identificar, ou seja, perceber a presença de uma entidade, e de Delimitar, ou seja, reconhecer as suas fronteiras. Um esquema é mostrado na Figura 11.

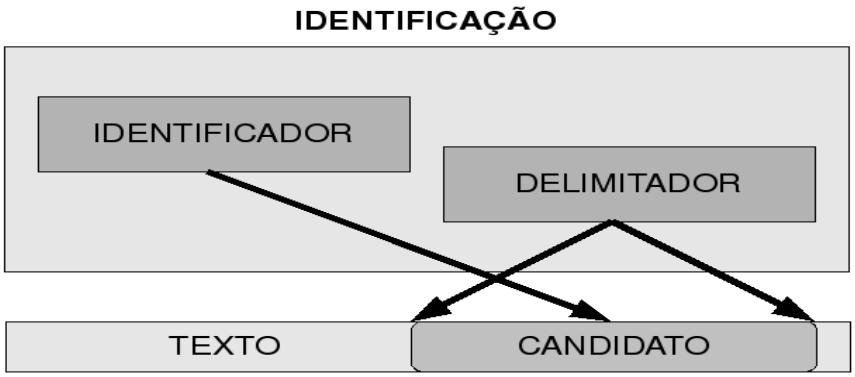

Figura 11: Arquitetura de Identificação

Nesta figura podemos ver a arquitetura de identificação dividida em dois módulos, o da identificação em si e o da delimitação da tarefa de identificar.

O processo de Classificação resume-se a um conjunto de regras de análise que determinam se um candidato pode ser considerado uma Entidade Mencionada, se seus limites foram identificados corretamente e, também, qual sua classe resultante.

Essas regras podem ser codificadas usando-se expressões regulares (ER). No Siemês (SARMENTO, 2006a) são descritas cinco regras usadas no processo de classificação. No sistema Malinche (SOLORIO, 2007) é mostrada a técnica de aprendizado adotada, que substitui, de certa forma, o uso de regras.

Regras, na maioria dos casos, trabalham junto a outros recursos importantes. O principal deles são gazetteers (SARMENTO et. al., 2006b). Sua principal tarefa é de encontrar similaridades entre as entidades que estão sendo classificadas com aquelas que já foram classificadas corretamente (presentes no gazetteer), fornecendo assim dicas de classificação.

Regras internas às Entidades Mencionadas e regras externas podem ser usadas. As externas são usadas geralmente na desambiguação. Elas são citadas como Metapalavras e Adivinhação na Seção 2.1. Os módulos de classificação podem ser vistos na Figura 12.

Na Figura 12, podemos ver os módulos do processo de classificação. Revisão da Identificação, módulo que melhora a qualidade da identificação; Regras definem regras de acesso aos recursos lingüísticos, tais como gazetteeers e almanaques; e Desambiguação, que melhora a qualidade da classificação. Estas técnicas podem ser substituídas pelo aprendizado de máquina. Gazetteer, Almanaque e Metapalavras são usados como recursos lingüísticos e são independentes dos critérios de classificação. NADEAU (2006) cita que os recursos lingüísticos utilizados são tão importantes quanto o processo utilizado na classificação.

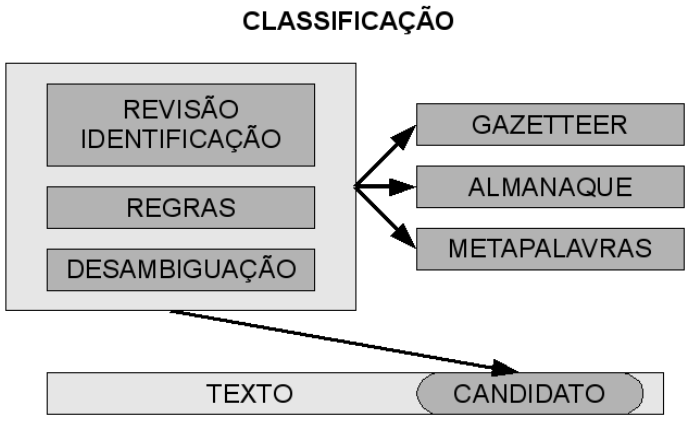

Figura 12: Módulos de Classificação

3.3.2 Levantamento de medidas avaliação atualmente usadas

Avaliação é um tópico muito importante no contexto do desenvolvimento de sistemas de REM. Seu resultado permite a comparação das técnicas utilizadas em diversas abordagens. Também, permite a avaliação entre sistemas diferentes, comparando qual obtém a melhor pontuação em uma determinada avaliação.

Diversas conferências definem medidas de avaliação, dentre elas: HAREM (português), MUC (inglês), IREX (inglês), CONLL (inglês) e ACE (inglês). Destas conferências exemplificaremos a usada no MUC, e por fim falaremos na adotada no HAREM. Para exemplificar, vamos mostrar um texto anotado por um humano, um texto anotado por um sistema e comparar ambos.

Texto original:

A Autoridade da Concorrência formalizou hoje a decisão de não oposição à compra da Carrefour Portugal pela Sonae Distribuição, que fica obrigada à alienação, de pelo menos, dois supermercados Modelo.

Texto anotado por humano:

A <INSTITUICAO>Autoridade da Concorrência</INSTITUICAO> formalizou <DATA>hoje</DATA> a decisão de não oposição à compra da <EMPRESA>Carrefour Portugal</EMPRESA> pela <EMPRESA>Sonae Distribuição</EMPRESA>, que fica obrigada à alienação, de pelo menos, dois supermercados <EMPRESA>Modelo</EMPRESA>.

Texto anotado por um sistema, exemplo:

A <EMPRESA>Autoridade da Concorrência</EMPRESA> formalizou hoje a decisão de não oposição à compra da Carrefour <PAIS>Portugal</PAIS> pela <EMPRESA>Sonae Distribuição</EMPRESA>, que fica obrigada à <CIENCIA>alienação<CIENCIA>, de pelo menos, dois <EMPRESA>supermercados Modelo</EMPRESA>.

Problemas de Identificação e Reconhecimento de Entidades podem ser vistos em Tabela 4.

Tabela 4: Comparação da anotação do sistema com a anotação correta

Delim.	Classif.	Anotação Correta	Anotação Sistema
correto	errado	<INSTITUICAO>Autoridade da Concorrência</INSTITUICAO>	<EMPRESA>Autoridade da Concorrência</EMPRESA>
-[13]	-	<DATA>hoje</DATA>	hoje
errado	errado	<EMPRESA>Carrefour Portugal</EMPRESA>	Carrefour <PAIS>Portugal</PAIS>
-[14]	-	alienação	<CIENCIA>alienação<CIENCIA>
errada	correto	<EMPRESA>Modelo</EMPRESA>	<EMPRESA>supermercados Modelo</EMPRESA>

Podemos ver que dois tipos de medidas de avaliação do sistema são consideradas: Classificação e Delimitação. Em relação a elas, podemos ter:

- Número de Respostas Corretas (NRC): Número de acertos do sistema (Tipo + Identificação)

- Número de Entidades Reais (NER): Número de entidades presentes na solução correta (Tipo + Identificação).

- Número de Entidades Identificadas (NEIS): É o número de Entidades Identificadas no Sistema (Tipo + Identificação). A pontuação final é dada pela medida-F.

No caso do exemplo acima temos os valores:

Precisão: NRC/NER = 4/10

Recall: NRC/NEIS = 4/10

Medida F = 40%

No HAREM também são adotadas as medidas de precisão, abrangência e Medida F, mas a forma de pontuar é diferente. Esta avaliação não considera um formato binário de avaliação (ou acerta ou erra). No HAREM há a possibilidade de acertar 30% na identificação, se identificarmos 30% de uma entidade, por exemplo. Detalhes da avaliação podem ser visto em SANTOS et. al. (2006c).

[13] Problema de não identificação de uma entidade
[14] Identificação de Entidade não existente

A primeira das medidas levantadas é a de precisão. Esta medida relaciona a quantidade de respostas corretas com a quantidade de respostas obtidas. Outra medida importante e complementar a esta primeira é a abrangência, que relaciona a quantidade de respostas corretas em relação ao número de todas as possibilidades de acerto. São medidas muito utilizadas em Recuperação de Informação. Outra medida que leva em consideração as duas acima é a medida F, dada abaixo:

Medida-F = 2*precisão*abrangência / (precisão + abrangência)

É a principal medida usada em campeonatos de avaliação de sistemas de REM, tal como o HAREM e o MUC.

3.3.3 Levantamento de Dificuldades/Limitações e Melhorias em sistemas de REM

Nesta seção são apontadas as principais dificuldades encontradas em sistemas de REM assim como formas de resolvê-las. Para este fim, usaremos o exemplo abaixo que pode ser visto em ARANHA (2007a):

1 – Fernando H. Cardoso

2 – Juiz Nicolau dos Santos Neto

3 – Presidente da Câmara dos Vereadores Alcides Barroso

4 – Hollywood

Em (1) encontramos uma dificuldade, que apesar de para um ser humano a identificação ser direta, para o computador se torna difícil, pois uma abreviação no meio de um nome pode ser entendida como fim de sentença. Especialmente se uma frase termina em Fernando H. e a próxima começa em Cardoso. Este problema pode ser parcialmente resolvido colocando-se uma regra especial, apenas para este caso específico. Como abreviações podem aparecer das formas mais imprevisíveis possíveis, necessitaremos, então, de várias regras deste gênero.

No exemplo (2) a presença de “dos” (letra inicial em minúscula) pode induzir um sistema de REM a separar os nomes ”Juiz Nicolau” e “Santos Neto” que, na verdade, fazem parte de um mesmo nome.

Em (3) há exatamente duas entidades. Mas não há seus limites, por exemplo, poderemos classificá-las erroneamente em Presidente da Câmara dos Vereadores e Alcides Barroso. Também não há nenhuma dica ortográfica que possibilite-nos a fazer a correta delimitação para esta entidade. Aliás, as dicas existentes confundem a delimitação correta.

Já em (4) temos um problema de ambigüidade: não podemos dizer se Hollywood significa lugar ou marca de cigarro. Na maioria dos casos, podemos usar o contexto para desambiguizar essas alternativas.

Outros pontos de dificuldade são inícios de frase, que é iniciado sempre por letra maiúscula por convenção. Logo, perdemos as dicas ortográficas na identificação. ARANHA (2007a) usa a estratégia de identificar a entidade se ela se enquadrar na classe dos substantivos. Tal dificuldade ocorre de forma similar para títulos e subtítulos.

3.4 Resultados Obtidos

Nesta seção, avaliaremos o protótipo quanto às medidas de precisão, abrangência e medida-F. O resultado para a tarefa de identificação é apresentado na Tabela 5:

Tabela 5: Resultado para a tarefa de identificação

Precisão	66.64
Abrangência	72.16
Medida F	69.29

O resultado para a tarefa de classificação em Tabela 6.

Tabela 6: Resultado para a tarefa de classificação

Precisão	44.53
Abrangência	48.53
Medida F	46.45

Os resultados obtidos são modestos, mas atingem a proposta inicial e têm possibilidade de melhorias se mais recursos lingüísticos forem adicionados. Um melhor resultado na tarefa de identificação (medida F de 69.29) em relação à classificação (medida F

de 46.45) é resultado de uma maior dificuldade da segunda tarefa, pois exige recursos lingüísticos e uma maior análise entre eles.

3.5 Dificuldades e Limitações

Alguns dos problemas encontrados são citados a seguir:

- Poucos recursos lingüísticos disponíveis. O único recurso lingüístico externo usado neste projeto é o Repentino. Não usamos a Coleção Dourada por ser um córpus muito pequeno. Outros corpora que possibilitariam uma melhor qualidade na desambiguação da tarefa ainda não existem ou não estão disponíveis.

- Complexidade de tempo: O gazetteer Repentino, que possui mais de 400 mil instâncias, é acessado a cada nova entidade encontrada. Logo, as implementações foram otimizadas para que não sofresse limitações de tempo de acesso. A busca seqüencial geraria 400 mil comparações enquanto a busca binária apenas 20.

4. Conclusão e Trabalhos Futuros

Este trabalho abordou o tópico de Reconhecimento de Entidades Mencionadas, resultando no desenvolvimento de um protótipo para a tarefa. Esta tarefa é importante na recuperação de informação em textos uma vez que nomeiam as entidades do mundo em questão de acordo com a ontologia adotada.

Avaliações estatísticas foram realizadas apenas para o levantamento da qualidade do protótipo desenvolvido. Futuramente, testes com usuários reais serão realizados. Isto possibilitará a verdadeira avaliação no aumento da inteligibilidade da aplicação. E, a continuidade deste trabalho no projeto de graduação II almejará obter um ganho ainda maior na inteligibilidade. Também, participação no campeonato HAREM é almejada. Isto possibilitará a comparação deste sistema com os sistemas REM atuais.

5. Comentários sobre o curso de graduação

O curso de graduação do ICMC é bem fundamentado. Fornece base técnica e teórica, que possibilita o desenvolvimento de projetos, dos mais variados tópicos, em computação.

Porém, seria interessante a inserção de uma matéria, ou talvez uma ênfase, na área de Processamento de Linguagem Natural. O ensino de técnicas específicas PLN, no nível da graduação, podem ser muito úteis aos futuros pesquisadores da área.

6. Referências Bibliográficas

ALUÍSIO, S.M.; NUNES, M.G.V.; PARDO, T.A.S., FORTES, R.P.M., PIMENTEL, M.G. (2007). "PorSimples: Simplificação Textual do Português para Inclusão e Acessibilidade Digital". Projeto do edital Fapesp-Microsoft Research, vigência: 11/2007 a 10/2009. Disponível em: http://caravelas.icmc.usp.br/wiki/index.php/Projeto_PorSimples.

ARANHA, C.N. (2007a). Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. Dissertação de Doutorado. PUC-Rio.

ARANHA, C.N. (2007b). O Cortex e a sua participação no HAREM. Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, Capítulo 9, p. 113–122.

BABYCH, B.; HARTLEY A. (2003). Improving Machine Translation quality with automatic Named Entity recognition. In: EACL 2003. 10th Conference of the European Chapter. Proc. Of the 7th Int. EAMT workshop on MT and other language technology tools. Budapest Hungary pp. 1-8.

BAKER, C.F.; FILLMORE, C.J.; LOWE, J.B. (1998). The Berkley FrameNet project. In the Proceedings of COLIN/ACL, pp. 86-90, Montreal.

BORBA. F.S. (1996). Uma gramática de valências para o Português. Editora Ática.

CARDOSO N.; SANTOS D. (2007a). Diretivas para a identificação e classificação semântica na coleção dourada do HAREM. Relatório Técnico DI/FCUL TR–06–18, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.

CARDOSO N.; SILVA M. J.; ANTUNES, M. (2007b). Validação estatística dos resultados do Primeiro HAREM. Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, Capítulo 5, p. 59–77.

CARDOSO N.; SANTOS D. (2007c). Reconhecimento de entidades mencionadas em português. Documentação e atas do HAREM, a primeira avaliação conjunta na área. Livro HAREM.

CARDOSO N.; SANTOS D. (2007d). Reconhecimento de entidades mencionadas em Português – Apêndice A. Documentação e atas do HAREM, a primeira avaliação conjunta na área. Livro HAREM.

CARRERAS X.; PADRÓ. L.; (2002). A flexible distributed architecture for natural language analyzers. Em Manuel González Rodrigues e Carmen Paz Suarez Araujo, editores, Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Espanha. 29-31 de Maio de 2002. p. 1813–1817.

CORTES, C.; VAPNIK V. (1995). Support-Vector Networks. Machine Learning. Springer Netherlands. Volume 20.

PEREIRA, J.P. (2007). Empresa brasileira desenvolve programa que aprende a ler textos. [http://www.cortex-intelligence.com/imagens/noticias/cortex_no_publicopt.pdf]

GRISHMAN R.; SUNDHEIM B. (1996). Message Understanding Conference - 6: A Brief History. In Proc. International Conference on Computational Linguistic.

KINGSBURY, P.; PALMER, M; (2002). From Treebank to PropBank. In the Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las palmas.

KIPPER, K.; DANG, H.T; PALMER, M. (2000). Class-based Construction of a Verb Lexicon. In the Proceedings of AAAI 17th National Conference on Language Resources on Artificial Intelligence. Austin, Texas.

LEFFA, V.J. (1996). Fatores da compreensão na leitura. Cadernos no IL, Porto Alegre, v.15, p.143-159, 1996. [http:// www.leffa.pro.br/fatores.htm]. Acesso em jan. 2008.

MARTINS. B.; SILVA. J.S.; CHAVES. M. S. (2006). O sistema CaGE no HAREM - reconhecimento de entidades geográficas em textos em língua portuguesa. Documentação e actas do HAREM, a primeira avaliação conjunta na área, Capítulo 8, p. 97–112.

MIHALCEA, R.; MOLDOVAN D. (2001). Document Indexing Using Names Entities, In: Studies in Informatics and Constrol, vol. 10, no. 1, January.

NADEAU D.; SEKINE S. (2006). A survey of named entity recognition and classification: in National Research Council Canada/New York University.

OLIVEIRA E.; XAVIER. M.A.; BAPTISTA J.; TRANCOSO I.; OLIVEIRA L.; MAMEDE N.; QUENTAL V.; NUNES G.; TEIXEIRA G.; RINO L.; VIEIRA R.; SARDINHA T. B.; FARIA I. H.; OLIVEIRA Jr. O. (2003). Pareceres sobre Linguateca: Relatório relativo ao período 2000-2003. [http://www.linguateca.pt/documentos/Pareceres.doc].

PARDO, T.A.S. (2002). DMSumm: Um Gerador Automático de Sumários. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos – SP.

PARDO, T.A.S.; RINO, L.H.M.; NUNES, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal. June 26-27.

PARDO, T.A.S. (2005). Métodos para análise discursiva automática: Dissertação de Doutorado. ICMC - USP. São Carlos - SP.

SANTOS, D.; CARDOSO, N. (2006a). A Golden Resource for Named Entity Recognition in Portuguese. Linguateca: Node of Oslo at SINTEF ICT. Linguateca: Node of XLDB at University of Lisbon.

SANTOS, D.; SECO, N.; CARDOSO, N.; VILELA, R. (2006b). HAREM: An advanced NER evaluation contest for portuguese. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genova, Italy.

SARMENTO. L. (2006a). A Named-Entity Recognizer for Portuguese Relying on Similarity Rules. PROPOR 2006.

SARMENTO, L.; PINTO, A.S.; CABRAL, L. (2006b). REPENTINO: A wide-scope gazetter for Entity Recognition in Portuguese – PROPOR´2006.

SOLORIO, T. (2007). MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish. Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, Capítulo 10, p. 123–136.

TORRES, E.F.; MAZZONI, A.A.; ALVES, J.B.M. (2002). A acessibilidade à informação no espaço digital. Ci. Inf. vol.31 no. 3. Brasília Sept./Dec.

VAPNIK V.N. (1995). The Nature of Statistical Learning Theory. Springer. Nova Iorque, NY, EUA.

Apêndice A. Projeto de Graduação II

O projeto de graduação II tratará da identificação e explicitação da estrutura argumental de verbos em textos em português. Uma explicação do problema é dada abaixo.

A.1 Identificação da Estrutura Argumental (IEA) dos Verbos com objetivo de aumento da Inteligibilidade em textos do Português

Na obra de Borba (1996) é citado um tipo de gramática, denominada de gramática de valências, visando tratar o fenômeno de palavras que não são auto-suficientes semanticamente. Valência é o nome dado ao número de argumentos necessários para que o entendimento destas palavras. No escopo do projeto PorSimples, somente verbos serão trabalhados. Em suma, serão realizadas a identificação de verbos e sua respectiva estrutura argumental. Podemos ver a aplicação deste conceito em uma ferramenta on-line da empresa *Cortex Intelligence*. Nesta ferramenta é disponibilizada uma série de textos e funções para recuperação das entidades mencionadas, dos verbos finitos[15], infinitos e alguns substantivos dos textos respectivos. Como exemplo, utilizaremos o texto da Figura 13.

A americana Amazon , maior varejista on-line , anunciou ontem que seu lucro líquido caiu para US$ 19 milhões , em comparação aos US$ 30 milhões do terceiro trimestre de 2005 . Ainda assim , o resultado bateu as expectativas do mercado e as ações subiram 12% . O faturamento da companhia teve aumento de 24% , para US$ 2,3 bilhões .

IDV sob nova direção

Figura 13: Texto exemplo disponibilizado no site da empresa *Cortex Intelligence*

[15] A flexão do verbo é finita quando traz informação de tempo e modo e infinita quando indeterminada em tempo e modo. Exemplo de flexões finitas: fizemos, fazíamos e faremos. Exemplos de infinitas: fazer, fazendo e feito.

É disponibilizada uma opção, via o botão “Ações” (veja Figura 1), que recupera todos os verbos no texto em foco (são apresentados em verde) e alguns substantivos (são apresentados em rosa). Para o texto da Figura 13 obtivemos o seguinte resultado, mostrado na Figura 14.

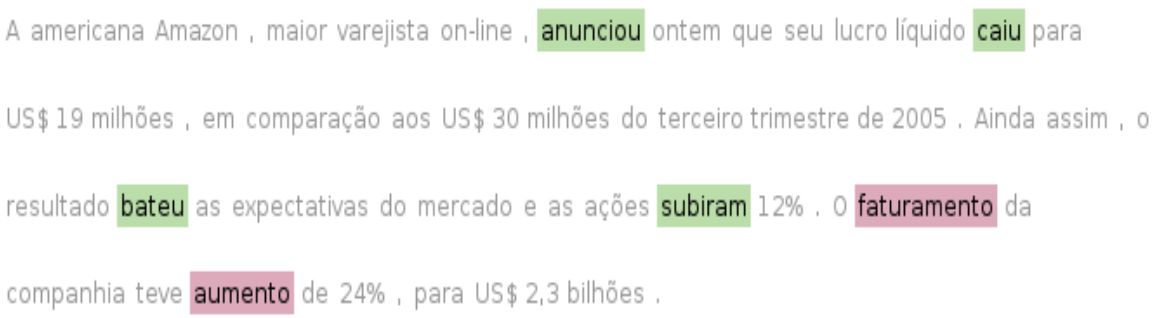

Figura 14: Ações Identificadas

Cada verbo (ou substantivo) identificado realiza uma função no contexto em que se encontra. Por exemplo, o verbo “anunciou”, na primeira linha da Figura 14, pressupõe que alguém anuncia (no caso a “Amazon”), e esse alguém anuncia algo (sua queda de lucro líquido). Já o substantivo “aumento” pressupõe que algo teve um aumento (“faturamento”) em uma determinada quantidade ou proporção, como mostrado na Figura 15.

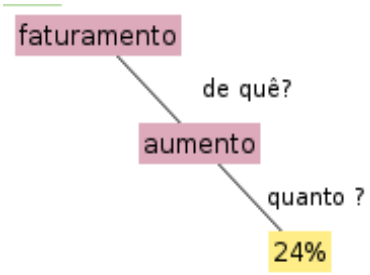

Figura 15: Estrutura argumental do substantivo “aumento”

Em Pardo (2005), podemos encontrar o uso do *Noisy-Channel*, um modelo de aprendizado de máquina (estatístico) para a identificação da estrutura argumental dos verbos para o Inglês. Neste trabalho, foram citados três dos projetos mais conhecidos na área, para a língua inglesa: FrameNet (Baker et al., 1998), VerbNet (Kipper et al., 2000) e o PropBank

(Kingsbury et. al., 2002). Já para o Português é proposto o desenvolvimento de um sistema, além da apresentação de uma variedade de metodologias, desde as manuais até as automáticas, para a tarefa de IEA.

A avaliação do sistema a ser construído neste projeto não seguirá a forma tal como encontramos em Pardo (2005), mas terá uma avaliação com o usuário em relação à melhoria da inteligibilidade dos textos lidos.

Anexo A. Classes do Repentino

Neste anexo, são mostradas as 11 classes semânticas em que o gazetteer Repentino se baseia:

1. Classe Localização

Define entidades que nomeiam entidades que tenham um posicionamento geográfico no universo. 16 subtipos: Terrestre; Hidro; Espacial; Endereço; Socio-cultural; Religioso; Endereço Alargado; País/Estado; Povoação/Região/Div. Administrativa; Civil/Administração/Militar; Património/Monumento; Propriedade; Mitológico/Ficcional; Comercial/Industrial/Financeiro; e Infraestrutura.

2. Classe Organização

Define entidades compostas por uma ou mais pessoas que operam como parte de um todo. Organizações geralmente possuem objetivos, regras e estrutura interna. São divididos em 11 subtipos: Civil-Militar; Clubes; Desportiva; Empresa; Ensino/I&D; Governamental/Administrativa; Grupos de Interesse; Religiosa; e Socio-Cultural**.**

3. Classe Seres

Define entidades de seres reais, fictícios ou mitológicos. Grupos de pessoas que não formem organizações também fazem parte desta categoria. São divididos em 6 subtipos: Colectivo Humano; Geopolítico/Étnico/Ideológico; Humano; Mitológico; e Não-Humano.

4. Classe Evento

Define entidades que simbolizam nomes de eventos. São divididos em 8 subtipos: Desportivo; Socio-Cultural; Efeméride; Científico; Cíclico; Político; e Prémio/Galardão**.**

5. Classe Produtos

Define entidades de nomes de produtos. Representa produtos desde os industriais até os artesanais. São divididos em 15 subtipos: Ferramentas/Instrumentos; Consumíveis; Electrónica/Electrodomésticos; Financeiro; Formato; Gastronomia; Inspecção/Exame; Médico/Farmacêutico; Marcas; Serviços e Recursos; Sistemas Informáticos e Aplicações; Tarefa Manual/Artesanato; Vestuário/Utilidades; e Veículos.

6. Classe Arte/Mídia/Comunicação

Define entidades de objetos relacionados a Arte, Mídia e Comunicação. São divididos em 9 subtipos: Filme; Livro; Música; Multimédia; Periódico; TV/Radio/Teatro; Arte & Design; e Texto Académico/Científico.

7. Classe Papeladas

Define entidades relacionadas a Leis, Decretos, Tratados, ou seja, qualquer tipo de documentos que tem nome. São divididos em 8 subtipos: Lei; Acordo; Norma; Certificações; Impostos/Emolumentos; Planos e Procedimentos; e Documentos.

8. Classe Substâncias

Define entidades que incluem elementos, substâncias, minerais. São divididos em 4 subtipos: Grupo; Minério; e Substância.

9. Classe Abstração

Define entidades relacionadas a conceitos abstratos tal como disciplinas, ciências, etc. São divididos em 12 subtipos: Estado/Condição; Disciplina/Arte & Ofício; Período/Movimento/Tendência; Formulação Mental; Era/Época; Processo; Símbolo; Índice/Taxa; e Tipo/Classe.

10. Classe Natureza

Define entidades da natureza, tal como animais e vegetais. São divididos em 5 subtipos: Animal; Fisiologia; Micro-organismos; Vegetal; Fenómenos Naturais.

11. Outros

Define entidades que não se enquadram nas anteriores. Tipo único: Outros.