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

Avaliação de Métodos de Extração Automática de Termos para a Construção de Ontologias

Claudia Zavaglia
Leandro Henrique Mendonça de Oliveira
Maria das Graças Volpe Nunes
Maria Fernanda Teline
Sandra Maria Aluisio

NILC-TR-05-01

Janeiro 2005

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

Avaliação de métodos de Extração Automática de Termos para a Construção de Ontologias[1]

Claudia ZAVAGLIA –UNESP/IBILCE/Brasil - zavaglia@lem.ibilce.unesp.br
Leandro Henrique Mendonça de OLIVEIRA –NILC[2]--ICMC-USP/Brasil - leandroh@nilc.icmc.usp.br
Maria das Graças Volpe NUNES – NILC-ICMC-USP/Brasil - gracan@icmc.usp.br
Maria Fernanda TELINE – NILC-ICMC-USP/Brasil – mteline@icmc.usp.br
Sandra Maria ALUÍSIO – NILC-ICMC-USP/Brasil - sandra@icmc.usp.br

RESUMO: *Para a sistematização do conjunto de informações terminológicas de um domínio, é fundamental o uso de ferramentas computacionais para a extração de termos. Este trabalho apresenta a avaliação de métodos de extração automática de termos (EAT) para a construção da OntoEco, uma ontologia da Ecologia. A avaliação é feita com uma lista de referência, usando as métricas de precisão e revocação.*

1. Introdução

Ontologias têm sido utilizadas para a representação de informações que veiculem um entendimento semântico comum de situações variadas do mundo real. Na Web, o uso de ontologias pode fornecer uma base de informações comum, bem como padronizada, englobando conceitos-chave que possam ser utilizados por serviços requisitados para cada situação particular. Em comércio eletrônico, por exemplo, o conjunto de informações oferecido pela ontologia pode ser utilizado para unificar e integrar definições de produtos oferecidos pelos mais variados pontos de venda, com um formato padrão e único. Além disso, as ontologias podem ser utilizadas em sistemas de recuperação da informação para melhorar a precisão e revocação dos documentos recuperados numa busca. Precisão é dada pela razão entre o número de respostas (documentos) corretas obtidas e o número total de respostas recuperadas. Revocação é dada pela razão entre o número de respostas corretas obtidas e o número total de respostas corretas possíveis.

Para a sistematização do conjunto de informações terminológicas de um domínio, é fundamental o uso de ferramentas computacionais para a extração de termos. Para o português

[1] Essa pesquisa foi financiada pelo CNPq (Conselho Nacional de Pesquisa) por um período de 6 meses em nível de pós-doutoramento realizado pela Profa. Dra. Claudia Zavaglia, junto ao NILC – USP/São Carlos – Brasil.
[2] Núcleo Interinstitucional de Lingüística Computacional (http://www.nilc.icmc.usp.br/nilc/index.html)

do Brasil, muitos projetos de construção de repertórios terminológicos ainda utilizam a abordagem manual a partir de córpus para a extração de termos, segundo o critério semântico. Ainda que este critério semântico seja adequado, a extração manual é lenta, sujeita à subjetividade e à omissão de termos relevantes.

2. Objetivos

Este trabalho apresenta a avaliação de métodos de extração automática de termos (EAT) para a construção de ontologias, em especial, para a construção da *OntoEco*, isto é, uma ontologia que prevê três subdomínios da Ecologia, a saber: *Ecologia de Ecossistemas* – EEc; *Ecologia de Populações* – EP; *Ecologia de Comunidades* – EC, onde os termos foram etiquetados manualmente com informações morfossintáticas e semânticas concernentes à Estrutura *Qualia* do Léxico Gerativo (LG) de Pustejovsky (1995), tendo sido utilizada para tal a ferramenta computacional Protégé-2000. A avaliação foi feita com uma lista de referência, usando as métricas de precisão e revocação.

3. Extração Automática de Termos (EAT): especificação

Extração de informação (EI) é o processo de identificar automaticamente tipos específicos de entidades, conceitos, relações ou eventos em textos livres e armazenar esta informação de uma forma estruturada (Yangarber and Grishman, 2000). Sistemas de EI são construídos para diferentes tarefas como, por exemplo, identificação e classificação de nomes próprios (Appelt and Israel, 1999), extração de eventos e relações típicas de um domínio de conhecimento (Yangarber and Grishman, 2000), extração de multipalavras (Smadja, 1991; Piao et al, 2003) e extração de terminologia (Oh et al, 2000; Bourigault, 1992; Daille, 1996).

O crescimento explosivo de dados do tipo texto disponíveis na *Web* e as vastas quantidades de novos materiais eletrônicos propiciam a criação de novos termos e alterações nos seus significados, principalmente, em áreas dinâmicas de pesquisa. Dado que o desenvolvimento de terminologias é um trabalho difícil quando realizado manualmente, lingüistas computacionais, lingüistas aplicados, tradutores, intérpretes, jornalistas científicos têm se interessado pela extração automática de terminologias de textos. A extração automática de terminologias (EAT) tem sido de grande interesse para todos os tipos de aplicações do Processamento de Línguas Naturais (PLN) que trabalham com domínios especializados e que, conseqüentemente, necessitam de um vocabulário especial.

O gargalo da EAT é a sua avaliação, pois exige a opinião de especialistas, sendo esse processo caro e demorado. Por outro lado, contar com recursos como glossários ou dicionários, isto é, com listas de referências, também traz seus riscos, uma vez que tais recursos são incompletos, dada a constante produção de novos termos.

No contexto de extração de terminologia de textos, *termos* são unidades lingüísticas, isto é, palavras ou combinações de palavras, designando conceitos ou entidades de um campo altamente especializado da atividade humana. Uma coleção de termos, relacionada com uma área de pesquisa (ou domínio) em particular, usualmente forma um sistema conceitual coerente conhecido como *terminologia* (Bolshakova, 2001). Termos compostos, que correspondem a duas ou mais unidades lexicais, são menos propensos a ambigüidade do que termos simples e aparecem em maior quantidade nos textos especializados, e são mais simples de se extrair. Termos compostos são os preferidos dos métodos de extração automática (Estopà Bagot, 1999).

A grande maioria dos documentos técnicos e artigos científicos contêm termos que são explicitamente ou implicitamente definidos pelos autores e então usados em seus textos. Em oposição aos termos de terminologia que estão fixos no dicionário, é importante que termos recém introduzidos sejam levados em conta para um processamento automático adequado de textos científicos e tecnológicos, pois tais textos apresentam grande quantidade de termos que estão em uso e que ainda não foram inseridos nos dicionários por terem sido introduzidos recentemente ou terem escopo local de aplicabilidade. Tais termos são denominados termos de autor. Em um aspecto diacrônico, não existe uma fronteira bem definida entre termos de dicionário e de autor. Usualmente, os termos têm origem como sendo de autor e, conforme esses vão sendo utilizados em vários textos de um dado campo, eles podem ser considerados como termos de dicionário por sua alta freqüência. As formas usadas para introduzir um termo de autor em um texto variam, resultando em três tipos diferentes de termos de autor (Bolshakova, 2001): a) termo é explicitamente definido; b) termo é indefinido (sua definição está ausente), mas é visualmente exposto; c) termo não é nem definido nem exposto, sendo então escondido. Estas três formas devem ser consideradas pelos métodos de extração automática. A última delas causa grande dificuldade para certos extratores, em razão de que os extratores geralmente utilizam padrões morfológicos e morfossintáticos para reconhecer e delimitar as unidades terminológicas, e o fato de tais padrões estruturais constituírem um filtro bastante permissivo para identificar as unidades terminológicas de um determinado

domínio impede que tais extratores delimitem todos os termos dos textos especializados. Dessa forma, se forem utilizados padrões referentes somente à forma da unidade, a maioria dos candidatos a termo apresentará delimitações errôneas. Por esta razão, os extratores também deveriam possuir conhecimento semântico a fim de detectar e delimitar automaticamente as unidades especializadas de forma mais exaustiva e precisa.

Todas as unidades léxicas têm uma freqüência associada correspondendo ao número de vezes que elas aparecem em um córpus. A partir desta informação, é possível saber se uma palavra pode ou não ser um termo. Ou seja, substantivos que aparecem mais de um certo número de vezes podem ser considerados termos candidatos; palavras de outras categorias devem ser mantidas a fim de completar o processamento de termos compostos. Existem, porém, estatísticas mais elaboradas para a seleção de candidatos a termos, por exemplo, Informação Mútua, Coeficiente *Log-Likelihood* (Daille, 1996) e Coeficiente *Dice*[3]. Sistemas de extração que usam medidas estatísticas são chamados de sistemas baseados em *estatística*. Outra abordagem encontrada na literatura é a *lingüística*, em que os sistemas detectam padrões recorrentes de unidades terminológicas complexas, tais como "substantivo–adjetivo" e "substantivo–preposição–substantivo", por exemplo; e a *híbrida*, em que os sistemas começam a detectar algumas estruturas lingüísticas básicas, tal como expressões nominais, e depois de os termos candidatos terem sido identificados, uma estatística relevante é usada para decidir se eles correspondem a um termo. O inverso também é possível, começando-se com uma lista de candidatos levantados estatisticamente, sendo que a informação lingüística, neste caso, é usada para filtrar termos válidos desta lista.

4. Metodologia e Desenvolvimento

4.1. O *CórpusEco*

Para o desenvolvimento desta pesquisa, elaboramos uma base de textos especiais, o *CórpusEco*, concernente ao subdomínio da Ecologia para o português do Brasil. Esse repertório textual conta hoje com 260.921 ocorrências e está armazenado em uma ferramenta computacional que administra grandes quantidades de dados, o *Folio Views 4.1*. Os textos foram extraídos de partes dos livros "A Economia da natureza" e "Ecologia", da editora Guanabara Koogan, e de revistas, presentes no Projeto Lácio-Web[4], sendo respectivamente

[3] http://www.d.umn.edu/~tpederse/Group01/bsp.txt
[4] http://www.nilc.icmc.usp.br/lacioweb/index.htm

dos gêneros instrucional e informativo. A lista de referência utilizada para a avaliação dos métodos foi confeccionada com 694 termos das partes dos livros acima extraídos com o critério semântico, além de dois glossários especializados, e mais 1105 termos do Dicionário On-line do Jornal do Meio Ambiente[5]. Após eliminação de termos duplicados e intersecção com o *CórpusEco*, a lista totalizou 520 termos, pois somente avaliamos com os métodos de extração os uni, bi e trigramas dentre os vários tamanhos de termos coletados inicialmente para formar a lista de referência. A extração dos termos para compor a ontologia *OntoEco* foi feita, primeiramente, de forma manual, utilizando-se o critério semântico no processo de extração. De fato, utilizamos a metodologia da onomasiologia, já que partimos do significado ou conceito de um item lexical para o seu significante, ou seja, a identificação da sua forma.

4.2. As abordagens da EAT

Em seguida, partimos para a extração automática dos candidatos a termos e avaliamos métodos simples de EAT para uni, bi e trigramas das três abordagens existentes: a *estatística*, a *lingüística* e a *híbrida*. Esses métodos, num total de quinze, foram desenvolvidos no projeto ExPorTer[6] que os avaliou para um córpus do gênero científico, diferindo dos gêneros tratados nesse artigo. Os métodos empregam recursos simples como: (a) uma *stoplist* para eliminar certos advérbios e palavras e expressões recorrentes em textos do gênero científico; (b) padrões sintáticos para os termos do domínio, por exemplo, *<substantivo adjetivo>*, *<substantivo preposição adjetivo>*, levantados após a aplicação de um etiquetador *Part-Of-Speech com precisão de 97%, desenvolvido no NILC*[7]**;** (c) uma *lista de expressões e palavras* características de definições, descrições, classificações como “definido(a)(s) como”, “caracterizado”, “chamado(a)(s)”, “significa”, entre outras que são concentradoras de termos. Para todos métodos implementados, após o cálculo das medidas para unigramas, bigramas e trigramas, as listas resultantes desse cálculo são tomadas como entrada para um outro programa, que realiza a intersecção das listas de unigramas, bigramas e trigramas com suas respectivas Listas de Referência para o cálculo da *Precisão e Revocação*, como mostrado nas próximas três seções.

[5] http://www.jornaldomeioambiente.com.br/
[6] http://www.nilc.icmc.usp.br/nilc/projects/termextract.htm
[7] http://www.nilc.icmc.usp.br/nilc/projects/mestradorachel.html

4.2.1. Método Estatístico

As medidas estatísticas utilizadas nesse trabalho são quatro: Freqüência, *Log-likelihood*, Informação Mútua e Coeficiente *Dice*, implementadas no pacote para a extração de n-gramas NSP[8] (N-gram Statistics Package), com objetivo de eleger a melhor medida estatística para a extração automática de unigramas, bigramas e trigramas. Entende-se como melhor medida estatística a que apresentar a maior *precisão*, embora também tenha sido calculada a *revocação*. Após a geração das listas de freqüência para unigramas, bigramas e trigramas, foram realizados os cálculos da informação mútua, do *log-likelihood* e do coeficiente *dice* para bigramas, que utilizam como entrada a lista de freqüência gerada para os bigramas do córpus. Em seguida, foram realizados os cálculos da informação mútua e do *log-likelihood* para trigramas, que utilizam como entrada a lista de freqüência gerada para os trigramas encontrados no córpus. Para unigramas somente foi realizado o cálculo da freqüência, pois é a única medida para unigramas disponível no pacote NSP.

O método estatístico usando a medida de *Freqüência* para unigramas teve seu corte estabelecido em 20, sendo a sua Precisão de 9,48% e Revocação de 34,27%. Para bigramas usando a medida de *Freqüência* o corte foi de 18, para a medida de *Informação Mútua* de 0,0097, para a *Log-likelihood* de 53,0782 e para o *Coeficiente Dice* de 0,1689, sendo que os quatro valores de Precisão foram 20,31% e os quatro de Revocação 14,44% Para os trigramas a medida de *Freqüência* utilizou corte de 18, para a *Informação Mútua* de 0,0066 e de *Log-likelihood* de 113, 2980, os três com Precisões iguais a 2,41% e Revocações iguais a 10,23%. Deve-se ressaltar que mantivemos os mesmos cortes resultantes de análises realizadas no projeto ExPorTer, embora o córpus lá utilizado tenha o dobro do tamanho do *CórpusEco*. Também mantivemos a mesma *stoplist*, embora os gêneros tratados fossem diferentes. Acreditamos que essas decisões podem ter afetado as precisões dos métodos estatísticos e híbridos.

4.2.2. Método Lingüístico

O método lingüístico implementado baseia-se em expressões lingüísticas e indicadores estruturais, bem como nos padrões morfossintáticos dos termos do domínio de Ecologia desenvolvidos no Projeto Bloc-Eco[9].

[8] http://www.d.umn.edu/~tpederse/nsp.html

Tabela 1: Padrões morfossintáticos

Padrões morfossintáticos utilizados[10]		
Para unigramas	**Para bigramas**	**Para trigramas**
n / np / adj / verb	n_adj / n_n / adj_n / adj_adj n_adv	n_prep_n / n_prep_np / n_n_adj / n_adj_adj n_prep_adj

Dessa maneira, o método lingüístico baseou-se tanto no trabalho de Heid et al (1996), no sentido de realizar um pré-processamento lingüístico no córpus utilizado e posteriormente a realização de consultas sobre o mesmo, quanto no trabalho de Klavans e Muresan (2000; 2001a; 2001b), no sentido de realizar uma busca por expressões lingüísticas e indicadores estruturais que introduzem definições e os termos definidos. Entretanto, o método lingüístico aqui utilizado não se assemelha totalmente ao método proposto por Heid et al (1996) em razão do córpus não ter sofrido o processo de lematização. Por outro lado, o método aqui implementado fugiu um pouco da proposta feita por Klavans e Muresan (2000; 2001a; 2001b), em razão de não terem sido realizadas buscas por padrões somente de expressões de definições, mas também de classificações, descrições e outras que concentram termos, além de não ter sido utilizado um módulo de análise gramatical, responsável por identificar definições introduzidas por fenômenos lingüísticos mais complexos, tais como anáforas e apostos. As precisões para uni, bi e trigramas foram 2,74%, 1,31% e 0,89% e as Revocações 89,18%, 62,22% e 82,95%.

4.2.3. Método Híbrido

Para a abordagem híbrida foi gerado um conjunto de orações do córpus, aqui chamado de subcórpus, que apresentassem as expressões lingüísticas definidas no método lingüístico, de maneira que cada oração é impressa no subcórpus somente uma vez, independentemente do número de expressões que pode apresentar. Este procedimento representou a "parte lingüística" do método híbrido. O subcórpus de saída, constituído pelas orações que apresentaram alguma expressão lingüística, é tomado como entrada para o pacote NSP, representando assim a parte estatística deste método. A freqüência, única medida estatística para unigramas encontrada no pacote NSP, foi calculada para os unigramas do subcórpus, utilizado-se o mesmo corte determinado na abordagem estatística, ou seja, 20, sendo a

[9] http://www.nilc.icmc.usp.br/nilc/projects/bloc-eco.htm

Precisão de 12,76% e a Revocação de 23,25% . O cálculo da *Freqüência* também foi efetuado para os bigramas e os trigramas do subcórpus, realizando o corte na *Freqüência* 18 tanto para bigramas quanto para trigramas, como estipulado na abordagem estatística, com Precisões 41,18% e de 18,75% e Revocações de 7,78% e 3,41%, respectivamente. A medida de Informação Mútua para o brigramas foi calculada sem corte, com Precisão de 1,68% e Revocação de 65%.

4.3. A *OntoEco*

A *OntoEco* foi implementada em uma ferramenta computacional, no caso, um editor de ontologias de livre acesso disponível no mercado, o Protégé-2000. Essa ferramenta foi desenvolvida para diferentes linguagens para a Web Semântica, entre as quais RDF e RDF Schema, que permitem a estruturação de informações de um domínio específico e possibilitam a comunicação, por meio de um vocabulário comum, entre agentes de software e páginas da Web. Seu modelo de conhecimento é representado por meio de *classes* (conceitos no domínio de discurso – constituem uma hierarquia taxonômica), *instâncias* dessas classes, *slots* (que descrevem as propriedades e atributos das classes e instâncias), *facetas* (que são restrições de informações, especificando informações adicionais sobre propriedades) e *axiomas* que especificam contrastes adicionais. Esse modelo é baseado em frames, usa a arquitetura de metaclasse, ou seja, um template que é usado para definir novas classes em uma ontologia, e possibilita a especificação de herança múltipla e de classes abstratas. Em sua implementação, a *OntoEco* encontra-se dividida em duas grandes classes: *CLASSES* e *LEXICAL_UNIT*. A classe *CLASSES* possui uma *META-CLASS* por meio da subclasse *STANDARD-CLASS* implementada como a subclasse *SEM_CLASS_BASE*, ou seja, a classe semântica base que definirá o padrão de configuração de todas as classes e subclasses que estiverem vinculadas a elas. O mesmo ocorre para a classe *LEXICAL_UNIT*, que possui uma *META-CLASS* por meio da subclasse *STANDARD-CLASS* implementada como a subclasse *LEXICAL_UNIT_BASE*, ou seja, a unidade lexical base que definirá o padrão de configuração de todas as classes e subclasses (itens ontológicos) que estiverem vinculadas a elas. A relação de hiponímia/hiperonímia, ou *é_um* (*is-a*), serviu para organizar diversos termos-conceito. De fato, todos os termos que fazem parte da ontologia possuem a relação *é_um*, como identificadora do *genus terminus* que a conceitua. À luz da Teoria do Léxico Gerativo, a

[10] n = nome; np = nome próprio; adj = adjetivo; verb = verbo; adv = advérbio

relação de hiperonímia corresponde às informações veiculadas pelo papel Formal da Estrutura *Qualia*. No Protégé-2000, essa relação está representada por classes e subclasses. Além disso, previmos um *frame :FORMAL* para cada classe e subclasse, quando for necessária a sua especificação para a recuperação do conceito veiculado pelas classes e subclasses Dessa forma, temos como subclasses da superclasse *CLASSES*: *INTERAÇÃO; POPULAÇÃO; COMUNIDADE; ECOSSISTEMA; ENERGIA;* entre outras. Após a distribuição dos itens lexicais na estrutura ontológica delineada, definimos e mapeamos as relações semânticas existentes entre eles presentes nos papéis da Estrutura *Qualia* (Pustejovsky, 1995). Cada unidade lexical terminológica ativa no campo da Ecologia foi delineada a partir de vários campos de valor, distribuídas em tabelas.

5. Resultados e Análise

Como os valores de precisão foram baixos para os quinze métodos calculados, vejamos o total de candidatos a termos extraídos para cada abordagem que são efetivamente unidades terminológicas nos 150 primeiros termos dos métodos com melhor precisão, para analisarmos se pelo menos os métodos conseguem distinguir os termos com melhores escores, isto é, os primeiros das listas de candidatos. No caso de empate para a precisão, a lista da freqüência foi escolhida:

Tabela 2: Resultados das análises

150 candidatos a termos analisados	Abordagem estatística	Abordagem Lingüística	Abordagem híbrida
unigramas	**42**	**21**	**45**
bigramas	**51**	**4**	**19**
trigramas	**10**	**1**	**3**
TOTAL	**103**	**26**	**67**

Foi possível observar, na abordagem estatística, que, em sua maioria, os *unigramas* que não são termos são substantivos flexionados ou não, verbos conjugados ou no infinitivo, adjetivos, advérbios, abreviações, letras soltas. Dos 150 candidatos a termos analisados em ordem decrescente, 42 candidatos verificaram-se como termos efetivos, cuja categoria gramatical freqüente foi a do substantivo. Já para os *bigramas*, obtivemos o melhor resultado de toda a extração, ou seja, 51 termos de combinatória gramatical “Substantivo + Adjetivo”.

Por sua vez, dos 150 *trigramas* levantados, 10 são efetivamente termos do tipo "Substantivo + Preposição + Substantivo".

Na abordagem lingüística, todos os *unigramas* efetivamente termos são de categoria substantiva, isto é, 21 deles. Os outros candidatos a termos extraídos são do tipo adjetivo, como "dinâmico", "mortos", "caídos" ou do tipo verbo flexionado e no infinitivo, como "mudando", "determinam", "dividem", "preferem" e "viver", "mascarar" respectivamente, ou ainda palavras formadas por hífen, tais como "focas-elefante", "besouros-de-farinha". Já para os *bigramas* que se caracterizaram de fato como termos, a combinatória gramatical foi a do tipo "Substantivo + adjetivo", no singular ou no plural. A extração aponta para outras combinatórias que não resultaram na efetivação de termos, tais como: "Pronome + substantivo" (quaisquer registros), "Adjetivo + Substantivo" (tremendos aumentos), "Advérbios + Substantivos" (muitas populações). Por sua vez, a extração de *trigramas* foi extremamente baixa: somente um candidato a termo caracterizou-se como unidade terminológica, cuja combinatória gramatical foi "Substantivo + Preposição + Substantivo". As outras combinatórias foram do tipo: "Substantivo + Adjetivo + Adjetivo" (troncos mortos caídos) e "Substantivo + Preposição + Substantivo" (indivíduos à densidade, noite por satélite, cidades de países, árvores das florestas), essa última com uma alta freqüência de ocorrência.

Na abordagem híbrida, obtivemos o melhor desempenho do extrator para a captura de *unigramas*, a saber: 45 termos, cuja categoria gramatical mais freqüente foi a de substantivo, flexionado ou não em número. As classes gramaticais dos outros candidatos a termos que não se efetivaram foram do tipo "abreviação" (fig), "verbal" (pode, podem), "adjetiva" (grande, maior), "pronominal" (tais). Para os *bigramas*, a combinatória de todos as unidades terminológicas foi "Substantivo + Adjetivo". Os candidatos a temos que não se efetivaram apontam para as combinatórias "Adjetivo + Substantivo" (novas espécies) e "Substantivo + Adjetivo"(populações naturais). Já os trigramas efetivamente termos são do tipo "Substantivo + Preposição + Substantivo" e os que não se caracterizaram como termos também, tais como *número de espécies*, *número de indivíduos*, *ponto de vista*.

6. Considerações finais

Em um primeiro momento, a extração automática de candidatos a termos alimentou o delineamento, propriamente dito, da estrutura arbórea da ontologia na medida em que nos

forneceu unidades terminológicas que se caracterizaram como classes ou subclasses, tais como: *população*, *comunidade*, *energia*, *área*, *ecologia*, *tabela de vida*, entre outros. Entretanto, a grande utilidade da extração automática foi, justamente, para a seleção das unidades lexicais para a ontologia, que podem ser implementadas tanto como subclasses quanto como instâncias. Dos resultados obtidos e das análises feitas, embora os métodos tenham baixas precisões, constatamos que a abordagem híbrida foi a que nos trouxe melhores resultados no que concerne à qualidade da extração, embora a quantidade de candidatos a termos extraídos tenha sido mediana, como as outras duas abordagens. De fato, para unigramas, a abordagem híbrida teve o melhor desempenho, ao passo que, numa análise geral, a abordagem estatística teve seus resultados superiores para bigramas e trigramas, sendo inferior somente para os unigramas em relação à abordagem híbrida. Já a abordagem lingüística carece de maiores detalhamentos ainda para que possamos obter resultados melhores em uma próxima tentativa de extração. Embora os valores dessa nossa avaliação não tenham sido altos, eles revelam que a aplicação de métodos automáticos para a captura de unidades lexicais, após serem refinados e especializados para os gêneros tratados assim como para o tamanho de córpus em uso, poderá auxiliar de maneira eficaz o trabalho de extração do terminólogo, dado que são capazes de identificar diversos itens lexicais que considerados efetivamente termos de forma bastante rápida. Em uma extração de termos com base no critério semântico, essa captura de unidades lexicais demandaria tempo e seria certamente muito mais lenta e subjetiva, se comparada à máquina.

Referências Bibliográficas

APPELT, D.; ISRAEL, D. Introduction to Information Extraction Technology. IJCAI´99 Tutorial. 1999. Disponível em: www.ai.sri.com/~appelt/ie-tutorial/.

BOLSHAKOVA, E. Recognition of Author's Scientific and Technical Terms. LNCS 2004, 2001 p. 281-90.

BOURIGAULT, D. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th International Conference on Computational Linguistics*, COLING 1992, 1992. p. 977-981.

DAILLE, B. *Combined approach for terminology extraction*: lexical statistics and linguistic filtering, PhD thesis, University of Paris 7, 1994.

ESTOPÀ BAGOT, R. Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada). Tese de Doutorado. Universidade Pompeu Fabra, 1999.

HEID, U.; JAUβ, S.; KRÜGER, K.; HOHMANN, A. Term extraction with standard tools for córpus exploration. In: *4th International Congress on Terminology and Knowledge Engineering*, Wien. August, 1996.

KLAVANS, J. L.; MURESAN, S. DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from Online Text. In: *Proceedings of AMIA*, 2000.

KLAVANS, J. L.; MURESAN, S. Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text. In: *Proceedings of JCDL*, 2001a.

KLAVANS, J. L.; MURESAN, S. Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. In: *Proceedings of AMIA*, 2001 b.

PIAO, S. L.; RAYSON, P.; ARCHER, D.; WILSON, A.; MCENERY, T. Extracting multiword expressions with a semantic tagger. In: *Proceedings of the Workshop on Multiword Expressions*: Analysis, Acquisition and Treatment, at ACL'03, the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003. pp. 49-56.

PUSTEJOVSKY, J. *The Generative Lexicon.* Cambridge: The MIT Press, 1995.

SMADJA, F. *Retrieving Collocational Knowledge from Textual Corpora. An application*: Language Generation. PhD Thesis, Computer Science Department, Columbia University, 1991.

YANGARBER, R.; GRISHMAN, R. *Extraction Pattern Discovery through Corpus Analysis*. TR- 00-143, The Proteus Project, New York University. In: Proceedings of the Workshop Information Extraction meets Corpus Linguistics, Second International Conference on Language Resources and evaluation (LREC 2000), Athens, Greece, 2000.