Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Análise Comparativa de Métodos de Extração de Termos: Abordagens Linguística e Estatística

Lucelene Lopes, Leandro Henrique M. de Oliveira, Renata Vieira

Relatório Técnico N^0 053

Porto Alegre, Agosto de 2009

Análise Comparativa de Métodos de Extração de Termos: Abordagens Linguística e Estatística

Lucelene Lopes[1], **Leandro Henrique M. de Oliveira** [2], **Renata Vieira** [1]

[1]Faculdade de Informática – FACIN – PUCRS
Porto Alegre – RS – Brasil

lucelene.lopes@pucrs.br, renata.vieira@pucrs.br

[2]Instituto de Ciências Matemáticas e de Computação – ICMC – USP
São Carlos – SP – Brasil
Embrapa Informática Agropecuária – CNPTIA
Campinas – SP – Brasil

leandro@cnptia.embrapa.br,leandroh@icmc.usp.br

Abstract. *Este artigo apresenta uma comparação entre duas abordagens de extração automática de termos: as abordagens linguística e estatística. Cada uma destas abordagens foi aplicada através de ferramentas automáticas para extração a partir de* corpus. *Nos experimentos deste artigo foi utilizado um* corpus *da área de Pediatria em português e os termos extraídos foram comparados a uma lista de referência desenvolvida manualmente. A contribuição deste artigo reside na melhor compreensão dos métodos, além de uma análise comparativa das abordagens por métricas usuais da área, a saber: precisão, abrangência e* f-measure.

Abstract. *This paper presents a comparison between two different approaches of automatic term extraction: linguistic and statistical approaches. Each of these approaches was applied using* corpus *automatic extraction tools. The experiments were made over a* corpus *from Pediatrics in portuguese, and the extracted terms were compared with a hand made reference list. This paper contribution resides in a better comprehension of extraction methods, and a comparative analysis of their approaches by usual metrics: precision, recall and f-measure.*

1. Introdução

É clara a importância e a dificuldade da construção de ontologias para a estruturação, organização e disseminação de um conhecimento específico. Dentre as formas de construir ontologias, a construção a partir de textos é aquela que mais se presta a uma automatização e a tarefa de extração de termos é o ponto de partida para este processo [9]. Além disso, trata-se de uma etapa fundamental, pois, dela depende o sucesso de todas as demais etapas, uma vez que os termos extraídos devem ser a representação conceitual do domínio alvo.

Via de regra, os processos de extração automática de termos baseiam-se na análise de um conjunto de textos (*corpus*) do domínio de interesse [5]. As abordagens de extração

automática deste artigo situam-se neste campo de pesquisa. Especificamente, neste artigo é realizada uma análise comparativa entre duas abordagens de extração de termos (Liguística e Estatística) sobre um *corpus* específico da área de Pediatria, com o intuito de verificar qual delas é mais indicada e benéfica para um processo futuro de construção de ontologias.

É um consenso da área de processamento de linguagem natural que os métodos de extração de termos podem ser agrupados segundo a abordagem utilizada em: linguísticos e estatísticos. No entanto, esta divisão raramente é estanque, pois praticamente todos os métodos sempre tem ao menos algum componente de cada uma das abordagens. Métodos baseados em informações linguísticas sempre levam em consideração algum critério de frequência, assim como métodos baseados em informações estatísticas usualmente consideram algumas listas de palavras que seguem critérios linguísticos (*stoplist*). Desta forma, a quase totalidade dos métodos poderiam ser vistos como híbridos, porém para fins de classificação a área denomina de métodos linguísticos aqueles que tem a maior parte das decisões baseadas neste tipo de informação e, analogamente, denomina-se métodos estatísticos aqueles em que não se considera explicitamente informações linguísticas.

O método baseado em informações linguísticas utiliza uma ferramenta automática que parte de um *corpus* anotado sintaticamente e extrai os termos utilizando uma análise baseada na busca dos sintagmas nominais mais frequentes. Neste sentido, este método é semelhante ao trabalho de Bourigault *et al.* [4] que também extrai sintagmas nominais levando em consideração as categorias morfossintáticas e as principais relações sintáticas como por exemplo, sujeito, objeto direto e complemento proposicional (de nome, de verbo e de adjetivo). No entanto, o trabalho de Bourigault e seus colaboradores está baseado em uma ferramenta desenvolvida para a extração de termos sobre um *corpus* composto de textos em língua francesa.

O segundo método utilizado neste artigo segue uma abordagem claramente estatística onde os termos são extraídos com uma análise da frequência destes termos no *corpus*, posto que eles não estejam em uma lista prévia de termos (*stoplist*). Neste sentido, este método é semelhante aos trabalhos de Aubin e Hamon [1] e Fortuna *et al.* [7] que utilizam ferramentas específicas que estatisticamente analisam *corpus* com o propósito de inferir conceitos (extração de termos), e também tentam inferir uma hierarquia entre os termos extraídos.

A seção 2 descreve o *corpus* de Pediatria utilizado e o processo manual para obtenção da lista de termos que foi considerada como referência. As seções 3 e 4 apresentam as duas abordagens utilizadas para extração automática de termos. A seção 5 expõe os resultados obtidos e a comparação entre as duas abordagens apresentadas. Por fim, a conclusão sumariza a contribuição e sugere trabalhos futuros.

2. *Corpus* e Lista de Referência

O *corpus* utilizado nos experimentos é composto por 283 textos em português extraídos da revista bilíngüe Jornal de Pediatria (http://www.jped.com.br/), num total de 785.448 palavras. No entanto, para analisar a eficiência do processo, é necessário que haja uma lista de termos de referência.

A lista de termos de referência foi construída pelo projetos TEXTQUIM-TEXTECC da Universidade Federal do Rio Grande do Sul

(www.ufrgs.br/textecc). O objetivo da seleção e listagem desses termos do *corpus* de Pediatria foi a elaboração de dois glossários para apoio aos estudantes de tradução.Para a identificação dos itens desses glossários, o grupo realizou uma extração de n-gramas de textos puros (sem anotação lingüística) do *corpus* de Pediatria. Nesse processo foram aproveitados apenas os n-gramas com mais de 4 ocorrências no *corpus*. A partir de uma lista de 36.741 n-gramas, partiu-se para um processo de filtragem baseado em heurísticas que resultou em uma lista com 3.645 n-gramas considerados possivelmente relevantes para integrar os glossários.

Uma etapa posterior às heurísticas foi a avaliação manual sobre a relevância dos termos realizada por 5 estudantes de tradução com algum conhecimento do domínio. Esse resultado foi novamente refinado por meio de uma verificação manual com o objetivo de tornar a referência mais adequada ao propósito de criação de uma ontologia (definição de conceitos), visto que o objetivo inicial era a construção de glossários para estudantes de tradução. Finalmente obteve-se uma lista com 2.150 termos, sendo 1420 bigramas, 730 trigramas. Termos de composição maior que 3 palavras não foram considerados no presente trabalho.

3. Abordagem Linguística – EχATO$_{LP}$

Nesta abordagem o processo de extração de termos inicia-se com anotação linguística dos textos que compõem o *corpus* que é feita pelo *parser* PALAVRAS [3]. O parser PALAVRAS faz análise sintática através da construção de uma árvore na qual os nós terminais (folhas da árvore) são as palavras do texto e os não terminais representam as categorias da estrutura da frase. Os diversos textos entram como arquivos ASCII (txt) e o PALAVRAS tem na saída as informações representadas em um arquivo no formato XML. Este XML contém todas as frases devidamente anotadas linguisticamente, ou seja, cada uma de suas palavras é anotada conforme sua função sintática, semântica e suas características morfológicas.

Desse conjunto de arquivos XML que representa o *corpus* anotado são extraidos os Sintagmas Nominais (SN). Ao contrário das palavras isoladas cujo significado depende fortemente do contexto, quando SN são extraídos de um texto seus significados permanecem os mesmos [8]. Os SN podem ser classificados de acordo com o número de palavras (tokens) que o compõem. Neste artigo a análise de extração foi feita apenas sobre SN com 2 (bigramas) e 3 tokens (trigramas). Para extração automática uma ferramenta, chamada EχATO$_{LP}$, foi implementada para extrair SN anotados pelo *parser* PALAVRAS.

EχATO$_{LP}$ – Extrator Automático de Termos para Ontologias em Língua Portuguesa – é uma ferramenta que recebe um *corpus* anotado e extrai automaticamente todos os SN classificando-os segundo o número de tokens. A ferramenta porém utiliza um conjunto de heurísticas para refinar o processo de extração. Estas heurísticas tem base linguística com o propósito de eliminar ou refinar SN que não sirvam como possíveis conceitos de uma ontologia, especificamente:

- são eliminados SN que terminam com preposição, *e.g.*, *"criança acrescida de"*, *"dosagem diária para"*;
- são eliminados SN que possuem números, *e.g.*, *"década de 50","dois estudos"*;
- são excluídos os SN cujo o núcleo não for substantivo, nem nome próprio, nem adjetivo, *e.g.*, particípio passado *"valor superestimado"*, *"observado por outros"*;

- são excluídos os SN que iniciam com pronomes, *e.g.*, *"estas condições" "todas as crianças" "seus acompanhantes", "esses dados"*.
- são aceitos apenas sintagmas que possuem letras (acentuadas ou não) ou hífen, ou seja, SN que contém caracteres especiais são eliminados, *e.g.*, *"fator RH+"*, *"dupla mãe/neonato"*;
- SN que começam com artigos são armazenados sem a primeira palavra (o artigo), *e.g.*, *"a cicatriz renal"* é armazenado apenas como *"cicatriz renal"*;
- SN que terminam com conjunções (*e* e *ou*) são armazenados sem a conjunção, *e.g.*, *"baixo peso e"* e *"leite materno ou"* são armazenados, respectivamente como *"baixo peso"* e *"leite materno"*.

Os sintagmas extraídos podem ser compostos de um número qualquer de tokens, inclusive sendo apenas um unigrama. Na prática, a ferramenta agrupa os sintagmas extraídos em dez listas que contém respectivamente os sintagmas compostos por 1 a 9 palavras e a última lista contém sintagmas compostos por 10 ou mais palavras.

A ferramenta EχATO$_{LP}$ gera cada uma destas dez listas de termos em ordem decrescente de frequência no *corpus*. Desta forma, estas listas podem ser facilmente submetidas a pontos de corte que levam em consideração a frequência relativa ou absoluta, ou simplesmente serem usadas na sua totalidade.

Na extração de termos feita neste artigo foram considerados apenas SN que tiveram frequência absoluta igual ou superior a 4 ocorrências no *corpus*, ou seja, SN que aparecem 3, 2 ou apenas 1 vez não foram incluídos na lista final de termos extraídos. Esta seleção de termos insere no processo de extração, que segue claramente uma abordagem linguística, um componente estatístico conforme foi citado anteriormente.

4. Abordagem Estatística – NSP

A ferramenta NSP – *Ngrams Statistic Package* [2] é um conjunto de programas escritos na linguagem Perl desenvolvido para identificar e extrair n-gramas, uma sequência contínua de palavras (tokens). Atualmente na versão 1.09, o NSP (`www.d.umn.edu/ tpederse/nsp.html`) é utilizado principalmente para a extração e análise de n-gramas a partir de textos ou *corpus* textuais.

O processamento feito pela ferramenta NSP neste artigo utiliza apenas um dos programas da ferramenta, o programa `count.pl` que consiste em extrair um conjunto de termos com:

- o número de tokens especificado;
- uma lista de palavras que devem ser ignoradas durante o processamento (*stoplist*);
- um ponto de corte indicando um limiar inferior para o qual termos com uma frequência absoluta menor do que este limiar serão desprezados; e
- a regra de formação de tokens que define quais palavras serão aceitas.

O número de tokens identifica o tamanho dos termos que serão extraídos. Para este artigo foram extraídos do *corpus* bigramas e trigramas.

Um dos pontos centrais da utilização do NSP é a escolha do conjunto de palavras da *stoplist*, também conhecida como *stop words*. A definição destas palavras a serem desprezadas pode ser feita de forma compacta através de uma sintaxe própria da ferramenta. Usualmente são informados nesta lista palavras funcionais que apareciam com grande

frequência, tais como preposições, artigos, conjunções, e também uma quantidade significativa de advérbios que não apresentavam nenhum valor terminológico. Para minimizar esse problema, foi construída e aplicada uma *stoplist* com tais palavras, a fim de obter listas menores, apresentando termos com maior probabilidade de serem conceitos de uma ontologia. Dessa maneira, a *stoplist* aplicada neste experimento continha preposições, artigos, conjunções, advérbios e algumas palavras de demarcação estrutural do texto, como por exemplo: "Introdução", "Referências", "Bibliografia". Esta definição de *stop words* insere neste processo de extração, que permanece claramente estatístico, um componente linguístico, posto que classes sintáticas específicas devem ser inseridas na *stoplist*.

O ponto de corte informa ao programa quais valores de frequência absoluta de n-gramas devem ser desconsiderados durante o processamento. Usualmente, o ponto de corte é definido pelo tamanho do *corpus* [10]. Este cálculo é consenso no domínio da Linguística de *Corpus* e pode ser definido diretamente pela fórmula:

$$\text{Ponto de Corte} = (\text{tamanho do } \textit{corpus}/100.000) + 1$$

Este cálculo é baseado na premissa de que, em um determinado *corpus*, os candidatos a conceitos menos frequentes não possuem valor terminológico, visto que sua frequência é baixa, e, geralmente, em textos especializados os termos ocorrem com frequências maiores [6]. Para nossos experimentos o ponto de corte definido foi de 4 ocorrências dado o tamanho do *corpus* de Pediatria, que retiradas as palavras contidas na *stoplist*, é de 362.496 mil palavras. Logo, termos que foram encontrados 1, 2 ou 3 vezes apenas foram desprezados.

Já a regra de formação de tokens permite definir e especificar para o programa qual o padrão de tokens deve ser selecionado em uma determinada execução. Por exemplo, podemos especificar que os tokens desejados em um dado momento sejam aqueles que começam somente com letra maiúscula, ou somente os tokens que iniciam com vogais ou consoantes, ou Ngrams que contenham a preposição "de", ou ainda, a eliminação de caracteres irrelevantes para a análise, tais como aspas, pontuações e outros marcadores tipográficos. Um bom exemplo de uso da regra de formação de tokens para processamento de *corpus* em Português é a que inclui palavras acentuadas, já que a língua padrão do Pacote NSP é a língua inglesa, e as acentuações não são reconhecidas. Nos experimentos deste artigo, a regra de formação de tokens foi utilizada com este propósito, sendo então aceitos tokens compostos por letras maiúsculas, minúsculas e com os acentos usuais da língua portuguesa (áéíóúâêôàãõüç) mais o hífen (-) para considerar palavras compostas.

Além de aplicar as regras de formação de tokens, a *stoplist* e a definição do ponto de corte para extração dos candidatos a termos, houve também um pós-processamento de limpeza da lista de bigramas que retirou os candidatos a termos que eram originalmente substantivos próprios. Dessa forma, candidatos a termos como "São Paulo", "Porto Alegre" e "Sociedade Brasileira" foram também excluídos por não possuírem nenhum valor terminológico. Esta tarefa não utilizou nenhum conhecimento linguístico mais sofisticado, pois somente excluiu termos cujos tokens iniciavam com maiúsculas.

5. Experimentos

O *corpus* de Pediatria citado na seção 2 foi submetido a duas formas de extração descritas nas seções anteriores (seções 3 e 4). As listas de bigramas e trigramas extraídos

por ambas abordagens foram comparadas com uma lista de referência composta de 1420 bigramas e 730 trigramas. O primeiro resultado de cada uma das abordagens gerou listas compostas de 1248 bigramas e 608 trigramas para a abordagem linguística (EχATO$_{LP}$) e 3709 bigramas e 2550 trigramas para a abordagem estatística (NSP).

A comparação das listas extraídas (LE) com as listas de referência (LR) mostrou que a abordagem linguística encontrou 686 bigramas e 276 trigramas presentes nas listas de referência, ou seja, 686 bigramas e 276 trigramas na intersecção entre LE e LR. Analogamente, para a abordagem estatística a intersecção entre LE e LR foi de 1230 bigramas e 556 trigramas. Uma primeira análise superficial destes números parece indicar que a abordagem estatística é francamente melhor, pois a quantidade de termos extraídos é claramente superior. Porém, é necessário levar em conta não apenas o número de termos encontrados ($|LR \cap LE|$), mas também o tamanho de cada uma das listas extraídas ($|LE|$) e o tamanho da lista de referência ($|LR|$).

Com intuito de tornar objetiva esta comparação, foram utilizadas métricas quantitativas que expressam a precisão e a abrangência das listas obtidas, bem como o equilíbrio entre estes dois índices (*f-measure*). A precisão (P) indica a capacidade do método de identificar os termos corretos, considerando a lista de referência. Este índice é calculado pela primeira das fórmulas abaixo que é a razão entre o número de termos encontrados na lista de referência ($|LR|$) e na lista de termos extraídos ($|LE|$), ou seja, a cardinalidade da intersecção dos conjuntos LR e LE pelo total de termos extraídos (cardinalidade do conjunto LE). Analogamente, a abrangência (A) avalia a quantidade de termos corretos extraídos pelo método em relação ao tamanho da lista de referência. Finalmente, a *f-measure* (F) é simplesmente a média harmônica entre a precisão e abrangência.

$$P = \frac{|LR \cap LE|}{|LE|} \qquad A = \frac{|LR \cap LE|}{|LR|} \qquad F = \frac{2 \times P \times A}{P + A}$$

Para os experimentos realizados, os valores de precisão, abrangência e *f-measure* calculados estão indicados na última coluna da tabela 1.

Tabela 1. Número de termos encontrados para diversos pontos de corte

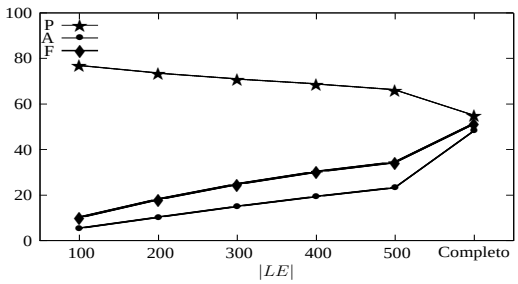
Métodos de Extração	Número de Termos	Tamanho da Lista					
		100	200	300	400	500	Completo
bigramas	$\|LE\|$	100	200	300	400	500	1248
EχATO$_{LP}$	$\|LR \cap LE\|$	77	147	213	275	331	686
bigramas	$\|LE\|$	100	200	300	400	500	3709
NSP	$\|LR \cap LE\|$	66	117	175	223	269	1230
trigramas	$\|LE\|$	100	200	300	400	500	608
EχATO$_{LP}$	$\|LR \cap LE\|$	48	97	151	206	236	276
trigramas	$\|LE\|$	100	200	300	400	500	2550
NSP	$\|LR \cap LE\|$	39	71	110	147	186	556

Estes resultados mostram que apenas a abrangência da abordagem estatística foi superior, sendo todos os demais índices favoráveis à abordagem linguística. Porém mais uma vez podemos estar sendo superficiais na análise dos resultados, pois não estamos levando em conta a distribuição de termos corretos (termos extraídos presentes em LR) nas listas extraídas em cada abordagem.

Desta forma, a tabela 1 apresenta o número de termos encontrados em cada uma das abordagens para diversos pontos de corte segundo a frequência dos termos. A última

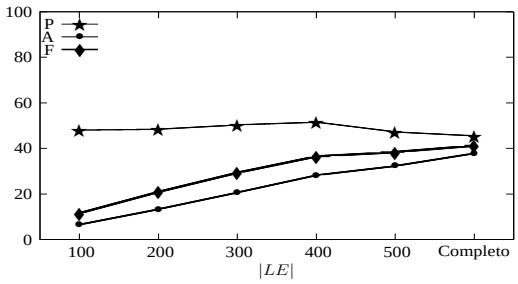
EχATO$_{LP}$ bigramas

$\|LE\|$	P	A	F
100	77,00%	5,42%	10,13%
200	73,50%	10,35%	18,15%
300	71,00%	15,00%	24,77%
400	68,75%	19,37%	30,22%
500	66,20%	23,31%	34,48%
1248	54,97%	48,31%	51,42%



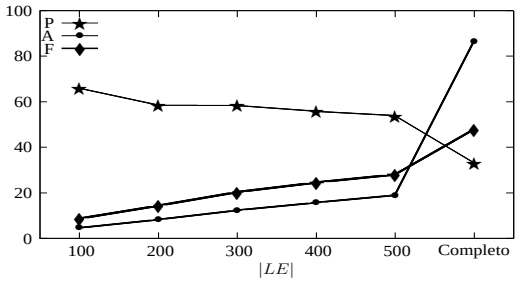
EχATO$_{LP}$ trigramas

$\|LE\|$	P	A	F
100	48,00%	6,58%	11,57%
200	48,50%	13,29%	20,86%
300	50,33%	20,68%	29,32%
400	51,50%	28,22%	36,46%
500	47,20%	32,33%	38,37%
608	45,39%	37,81%	41,26%



NSP bigramas

$\|LE\|$	P	A	F
100	66,00%	4,65%	8,68%
200	58,50%	8,24%	14,44%
300	58,33%	12,32%	20,35%
400	55,75%	15,70%	24,51%
500	53,80%	18,94%	28,02%
3709	33,16%	86,62%	47,96%



NSP trigramas

$\|LE\|$	P	A	F
100	39,00%	5,34%	9,40%
200	35,50%	9,73%	15,27%
300	36,67%	15,07%	21,36%
400	36,75%	20,14%	26,02%
500	37,20%	25,48%	30,24%
2550	21,80%	76,16%	33,90%

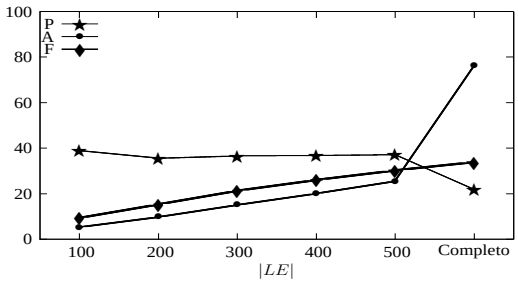


Figura 1. Métricas para listas reduzidas

coluna (Completo) representa os números já apresentados para as listas completas geradas pelas ferramentas EχATO$_{LP}$ e NSP. As demais colunas apresentam listas reduzidas por pontos de corte onde considera-se apenas os 100, 200, 300, 400 e 500 primeiros termos das listas extraídas, respectivamente. Igualmente, a figura 1 apresenta os valores e gráficos dos índices calculados para estes pontos de corte.

6. Conclusão

Foram realizados experimentos sobre um *corpus* de Pediatria em língua portuguesa. Sobre esse *corpus* listas de bigramas e trigramas foram extraídas através de duas abordagens diferentes, sendo uma delas predominantemente linguística e a outra fortemente estatística. Estas listas foram comparadas através das métricas de avaliação com uma lista de referência produzida manualmente sobre o mesmo *corpus*. Nessa comparação fica claro que a abordagem linguística sobressai a estatística, apesar dos resultados da abordagem estatística também serem consideráveis.

Ignorando as dificuldades da anotação linguística de um *corpus*, podemos concluir que a abordagem linguística utilizada pelo EχATO$_{LP}$ fornece melhores resultados e portanto é a mais indicada para o objetivo de identificação de conceitos na construção automática de ontologias. Cabe salientar que a dificuldade de anotação sintática consiste em ter uma ferramenta (*parser*) confiável e de fácil adaptação de sua saída para a ferramenta de extração linguística. Na verdade, alguns problemas encontrados nestes experimentos, como por exemplo, a menor abrangência do EχATO$_{LP}$ para bigramas e trigramas pode, talvez, ser explicado por erros de anotação herdados do *parser* PALAVRAS.

Já uma abordagem estatística como a utilizada pelo NSP tem a vantagem de ser um processo mais autocontido que implica em construir ou reaproveitar uma *stoplist* e um conjunto de regras de construção de tokens adequados. Definidas estas regras e lista, este processo pode ser generalizado até para outras línguas sem perda de generalidade.

A própria simplicidade da abordagem estatística contribui para que se possa identificar um grande número de termos. Este fato explica a grande abrangência desta abordagem que ocorre tanto para bigramas, quanto para trigramas. No entanto, esta mesma simplicidade que contribui para o aumento da abrangência custa caro ao reduzir em uma escala maior a precisão. Portanto, o aumento de abrangência não compensa a diminuição da precisão como pode ser observado pelos menores valores de *f-measure* quando comparado aos valores da abordagem linguística.

Em resumo, é possível afirmar que havendo confiança na ferramenta de anotação linguística e sua adaptação ao extrator de termos, é mais interessante usar uma abordagem linguística como a busca por sintagmas nominais apresentada neste artigo. Obviamente, esta afirmação esta baseada no estudo de um único *corpus* comparado a uma lista de referência desenvolvida manualmente. Portanto, um trabalho futuro natural ao descrito neste artigo é o estudo destas abordagens aplicadas a outros *corpus* para reforçar esta afirmação. Outra sequência natural é continuar o processo de construção de ontologias utilizando a lista de termos extraídos para identificar uma hierarquia entre eles.

Apesar disto, este artigo já apresenta uma contribuição significativa ao atribuir números a intuição de que um processo mais refinado de extração baseado em informações linguísticas supera abordagens mais simples. Isto se verifica claramente pelos índices indiscutivelmente superiores de precisão encontrados em todas as comparações, uma vez que para a construção de ontologias sempre será mais fácil lidar com poucos termos que sejam relevantes, do que com uma grande quantidade de termos sem relevância terminológica.

Referências

[1] AUBIN, S.; HAMON, T. Improving term extraction with terminological resources. FinTAL 2006, LNAI 4139, pp. 380-387, 2006.

[2] BANBERJEE, S.; PEDERSEN, T. The Design, Implementation, and Use of the Ngram Statistics Package. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Feb., 2003, Mexico City.

[3] BICK, E. The parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Arhus University, 2000.

[4] BOURIGAULT, D.; FABRE, C.; FRÉROT,C.; JACQUES,M.;OZDOWSKA, S. SYNTEX, analyseur syntaxique de corpus, TALN, Dourdan 2005.

[5] BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. In: P-Buitelaar, Cimiano, P.; and Magnini, B. (Ed.). Ontology Learning from Text: Methods, Evaluation and Applications, v. 123 of Frontiers in Artificial Intelligence and Apllications. IOS Press, 2005.

[6] ESTOPÁ 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.

[7] FORTUNA, B; LAVRAC, N.; VELARDI, P. Advancing topic ontology learning through term extraction. PRICAI 2008, LNAI 5351, pp. 626-635, 2008.

[8] KURAMOTO, H. Nominal Groups: a New Purpose to Information Retrieval. DataGramaZero - Revista de Ciência da Informação - v.3 n.1 Fev., 2002.

[9] LOPES, L.; VIEIRA, R.; FINATTO, M. J.; ZANETTE, A.; MARTINS, D.; RIBEIRO JR, L. C. Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. RECIIS, v.3, n.1, p.72-84, 2009

[10] TELINE, M. F. Avaliação de métodos para extração automática de terminologia de textos em português. Dissertação de Mestrado. ICMC/USP, 2004.