Hierarquias de conceitos extraídas automaticamente de *corpus* de domínio específico – Um experimento sobre um *corpus* de Pediatria

Lucelene Lopes, Renata Vieira, Daniel Martins

Grupo Processamento de Linguagem Natural (PLN), Faculdade de Informática (FACIN), Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Brasil

Resumo - Este artigo apresenta a extração automática de hierarquias de conceitos a partir de um *corpus* do domínio específico de Pediatria. O processo inicia na extração automática de termos simples e compostos. Especificamente, a extração de termos é feita, sob o ponto de vista linguístico, através da localização de sintagmas nominais requentes no *corpus*. A contribuição deste trabalho está em extrair e organizar de forma automática o conhecimento de um domínio específico em forma de hierarquias de conceitos, sendo essa uma forma de auxiliar diretamente a tomada de decisão na área da saúde utilizando a visualização de árvores hiperbólicas, ou ainda permitir a estruturação de ontologias.

Palavras-chave: Processamento de Linguagem Natural, Extração automática de termos, Hierarquias de conceitos, Ontologias

Abstract – This paper presents the automatic extraction of concepts hierarchy from a Pediatrics specific domain *corpus*. The process starts in the automatic extraction of simple and compound terms. Specifically, the term extraction is made, from a linguistic point of view, through frequent noun phrases identification. This paper contribution resides in the extraction and organization of a domain specific knowledge as concept hierarchies, to directly help decision making process in health care with hyperbolic trees visualization, or also to allow ontology construction.

Keywords Natural Language Processing, Automatic Term Extraction, Concepts Hierarchies, Ontologies

Introdução

O tratamento automatizado das informações textuais tende a auxiliar pesquisadores e gestores de políticas de informação em várias áreas do conhecimento. Muito se tem discutido sobre organização, compactação e representação da informação em ciências da saúde, no que diz respeito ao reconhecimento de linguagens e de terminologias científicas.

Para representar e organizar o conhecimento de um domínio específico de Pediatria, faz-se necessário a utilização de hierarquias de conceitos, que podem ser utilizadas tanto por sistemas computacionais como também por especialistas humanos. Para construção de hierarquias de conceitos uma etapa inicial de extração de termos relevantes do domínio em questão é fundamental (1). Esses termos podem ser utilizados para construção de hierarquias de conceitos, ontologias, bem como dicionários de termos. Todos estes possuem várias aplicações dentro das áreas de tradução automática, recuperação de informação, web semântica, etc.

Uma alternativa à consulta com um especialista para determinar os termos relevantes de um domínio específico é a utilização de processos automáticos de extração de termos onde a intervenção humana seja minimizada. Uma vez extraídos os termos, também é possível inferir uma hierarquia entre eles.

Neste sentido, este artigo apresenta um experimento de extração automática de conceitos e sua hierarquia através da extração de termos de um conjunto de textos (*corpus*) do domínio de Pediatria, tendo como base um trabalho anterior realizado sobre este mesmo *corpus* (2).

Este *corpus* foi anotado linguisticamente pelo *parser* PALAVRAS (3), e a ferramenta ExATOlp (4) foi utilizada para extrair listas de termos candidatos a conceitos. Cada termo possui uma frequência relativa associada que é

utilizada como critério de ordenação de relevância dos termos.
Os termos extraídos são comparados com uma lista de termos produzida manualmente por um grupo de linguístas (5) e métricas usuais da área de Processamento de Linguagem Natural foram usadas para avaliar essa comparação.
Após a avaliação do método automático de extração de termos, é necessário identificar dentre esses termos quais são os bons candidatos a conceitos do domínio, ou seja, aqueles que possuem uma semântica (significado) relevante ao domínio. Nesse experimento considera-se como conceito os termos que são ao mesmo tempo frequentes e linguisticamente relevantes.
Estabelecidos os conceitos, partiu-se para uma etapa de construção de hierarquia com o intuito de definir a base (hierarquia de conceitos) de uma ontologia sobre o domínio específico.

Métodos

O *corpus* utilizado nos experimentos possui 283 textos em português extraídos do Jornal de Pediatria (www.jped.com.br/) (6), num total de aproximadamente 785.448 palavras (tokens). Para analisar a eficiência do processo de extração é necessário uma lista de termos de referência.
A geração das listas de termos de referência utilizada nesse artigo foi feita numa etapa anterior, a partir de um trabalho original realizado pelo grupo de linguística da Universidade Federal do Rio Grande do Sul (TEXTQUIM-TEXTECC), que realizou um trabalho manual de extração de termos presentes no mesmo *corpus* de Pediatria visando à elaboração de um glossário para apoio aos estudantes de tradução (5).
A extração automática de termos propriamente dita foi feita em duas etapas que serão detalhadas a seguir: Anotação do c*orpus* pelo *parser* PALAVRAS; Extração de termos candidatos pelo ExATOlp.

Anotação do *Corpus*

A anotação linguística dos textos que compõem o *corpus* é feita pelo parser PALAVRAS. Os diversos textos entram como arquivos ASCII (txt) e tem como saída as informações representadas em um arquivo no formato TigerXML. Esse arquivo XML contém todas as frases devidamente anotadas linguisticamente, ou seja, cada uma de suas palavras são anotada conforme sua função sintática e suas características morfológicas e semânticas.
O parser PALAVRAS faz análise sintática utilizando-se da construção de uma árvore onde os nós terminais (folhas da árvore) são as palavras do texto e os nós não terminais representam as categorias da estrutura gramatical da frase.

Extração de Termos

A extração automática dos termos candidatos é feita pelo ExATOlp - Extrator Automático de Termos para Ontologias em Língua Portuguesa uma ferramenta que recebe um *corpus* anotado e extrai automaticamente todos os sintagmas nominais (SN) deste *corpus* classificando-os segundo o número de tokens.
SNs são importantes, pois, ao contrário das palavras isoladas cujo significado depende fortemente do contexto, os SNs guardam os seus mesmos significados quando são extraídos de um texto (7).
Os sintagmas extraídos são salvos em listas que podem conter tanto os SNs na sua forma original no texto, como em sua forma canônica (sem declinações e concordâncias de gênero grau e número). Neste trabalho utilizamos a forma original dos termos.
Foram extraídos sintagmas nominais constituídos por qualquer número de tokens. Apesar disto a ferramenta ExATOlp agrupa os termos extraídos segundo o número de tokens em unigramas (1 token), bigramas (2 tokens), trigramas (3 tokens) e assim por diante até termos com 10 ou mais tokens que são chamados de multigramas.
A análise de extração foi feita apenas sobre bigramas e trigramas, devido à necessidade de comparar os termos extraídos com a lista de referência (que contém apenas bigramas e trigramas).

Determinação de conceitos

As listas de bigramas e trigramas extraídos foram comparadas com uma lista de referência composta de 1420 bigramas e 730 trigramas. O resultado de extração automática gerou listas compostas de 1248 bigramas e 608 trigramas.
A comparação das listas extraídas (LE) com as listas de referência (LR) mostrou que a

abordagem linguística utilizada encontrou 686 bigramas e 276 trigramas presentes nas listas de referência, ou seja, 686 bigramas e 276 trigramas na intersecção entre LE e LR.
Com intuito de tornar numericamente 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 que é chamado de *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 o tamanho da lista de termos extraídos LE, ou seja, a cardinalidade da intersecção dos conjuntos LR e LE, pelo total de termos extraídos, isto é 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 medida harmônica entre a precisão e abrangência.

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

Com o objetivo de aumentar a confiança na extração de termos realizada, observou-se as métricas numéricas de qualidade acima para diversas versões limitadas das listas de termos extraídos. Especificamente, limitou-se a lista de termos extraídos a somente os 100, 200, 300, 400 e 500 termos mais frequentes, além da lista completa, ou seja, com todos os termos que apareceram pelo menos 5 vezes no *corpus*. Estas versões reduzidas das listas são denominadas na literatura (2) de "pontos de corte". O Quadro 1 apresenta o número de termos encontrados para diversos pontos de corte segundo a frequência dos termos.

Métodos de Extração	Número de Termos	Tamanho da Lista					
		100	200	300	400	500	Completo
bigramas	$\|LE\|$	100	200	300	400	500	1248
EχATOLP	$\|LR \cap LE\|$	77	147	213	275	331	686
trigramas	$\|LE\|$	100	200	300	400	500	608
EχATOLP	$\|LR \cap LE\|$	48	97	151	206	236	276

Quadro 1 – Número de termos encontrados para diversos pontos de corte

Neste quadro as primeiras colunas apresentam listas reduzidas por pontos de corte onde se consideram apenas os 100, 200, 300, 400 e 500 primeiros termos das listas completas extraídas. Igualmente, o Quadro 2 apresenta os índices calculados para estes pontos de corte.

EχATOLP bigramas				EχATOLP trigramas			
$\|LE\|$	P	A	F	$\|LE\|$	P	A	F
100	77,00%	5,42%	10,13%	100	48,00%	6,58%	11,57%
200	73,50%	10,35%	18,15%	200	48,50%	13,29%	20,86%
300	71,00%	15,00%	24,77%	300	50,33%	20,68%	29,32%
400	68,75%	19,37%	30,22%	400	51,50%	28,22%	36,46%
500	66,20%	23,31%	34,48%	500	47,20%	32,33%	38,37%
1248	54,97%	48,31%	51,42%	608	45,39%	37,81%	41,26%

Quadro 2 – Métricas para listas reduzidas

Com base nos experimentos verificou-se a qualidade de extração automática realizada em função dos altos índices de precisão obtidos para bigramas (55%) e trigramas (45%).
Confirmada a qualidade de extração, definiu-se como conceitos os termos extraídos com relevância semântica (SNs) que apareceram pelo menos 5 vezes no *corpus*.

Construção de Hierarquias

As hierarquias foram geradas através do desenvolvimento de um módulo que futuramente será integrado à ferramenta ExATOlp. Esse módulo organiza de forma hierárquica os termos extraídos através do núcleo do sintagma nominal, seja ele um substantivo, substantivo próprio, adjetivo ou verbo no particípio passado.
É importante salientar que esta abordagem de construção de hierarquia através do núcleo de sintagma nominal é uma contribuição original deste artigo, pois de acordo com o conhecimento dos autores, apenas abordagens léxico-sintáticas (8), de distribuição de similaridade (9), de análise de co-ocorrência, com o uso de dicionários prédefinidos (10), ou análise formal de conceitos – FCA (1) vem sendo empregadas.
As hierarquias foram geradas através de um módulo de software que tem como saída um arquivo OWL. Arquivos OWL podem ser utilizados como formato de entrada de vários sistemas com diferentes aplicações. Uma delas

é a disponibilização da hierarquia gerada para usuários do domínio específico.
Neste caso, devido a necessidade de interação com o profissional da área da saúde optou-se por utilizar a ferramenta TreeBolic (11), um visualizador de árvores hiperbólicas. Desta forma é possível a manipulação dos termos com árvores interativas que facilitam tanto uma visão geral de todos os nodos da hierarquia bem como uma visão mais restritiva de apenas parte deles.

Resultados

A hierarquia gerada para o *corpus* de Pediatria é composta de 2933 termos, sendo o primeiro nível composto por 720 unigramas e assim por diante até termos composto por 5 tokens.
No entanto, com o propósito de visualização através de árvores hiperbólicas um corte mais restritivo foi aplicado.
A Figura 1 apresenta uma pequena amostra da hierarquia obtida para o corpus de Pediatria com apenas os 9 termos mais frequentes do primeiro nível da hierarquia gerada. Cabe salientar que os termos em um mesmo nível da hierarquia são apresentados segundo suas ordens de frequência relativa no *corpus*.

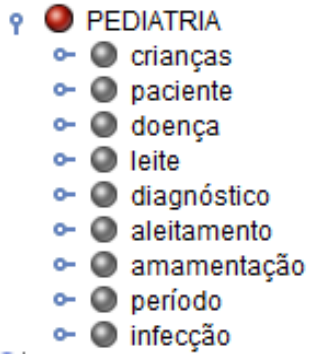

Figura 1 – Amostra da hierarquia de conceitos

A Figura 2 apresenta a mesma amostra da hierarquia gerada com um dos termos (infecção) expandido. Note-se que quando se expande um termo é possível verificar os seus níveis subalternos, como por exemplo, infecção viral e infecção viral de vias aéreas. Estes exemplos das Figuras 1 e 2 foram retirados de uma das opções da ferramenta Treebolic não em forma de árvore hiperbólica, mas sim como uma simples estrutura hierárquica de termos.

Figura 2 - Hierarquia expandida

A visualização de uma árvore hiperbólica é mais sofisticada como pode ser visto na Figura 3. É possível utilizar a visualização hiperbólica, onde em tempo real o usuário pode alterar a visualização da árvore de acordo com seus propósitos.

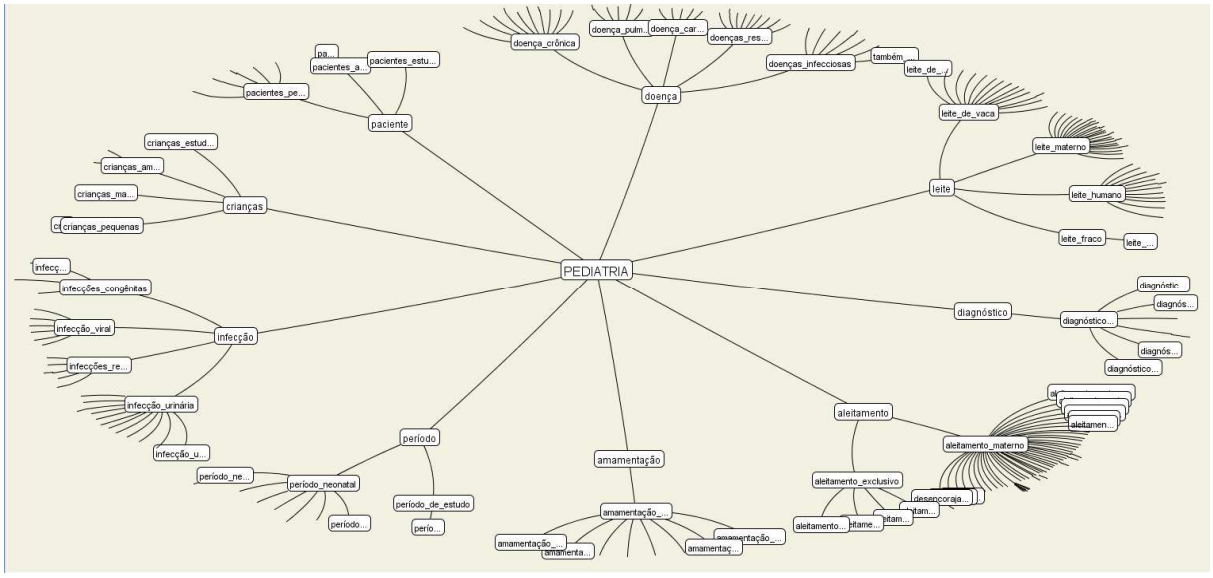

Figura 3 – Visualização Hiperbólica completa

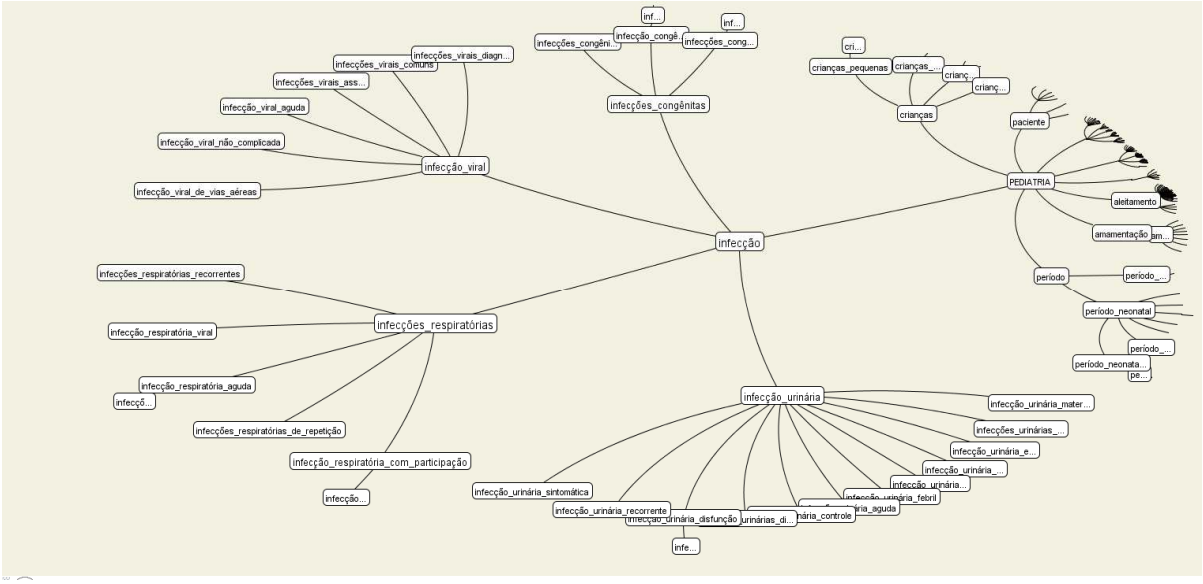

Figura 4 – Visualização hiperbólica de um nodo expandido da árvore

Um exemplo da mudança em tempo real da visualização é mostrado na Figura 4 que representa a mudança do foco central da árvore para um dos nodos do primeiro nível. Neste exemplo é possível visualizar o nodo "infecção" em maior detalhe, e assim analisar todos os possíveis termos relacionados a ele, ou seja, seus subníveis.

Conclusão

Foram feitos experimentos sobre um *corpus* de Pediatria em língua portuguesa. Sobre esse *corpus* listas de termos com diversos números de palavras foram extraídos através de uma abordagem predominantemente linguística. Dentre as listas extraídas, duas delas foram comparadas utilizando métricas de avaliação

com uma lista de referência produzida manualmente sobre o mesmo *corpus*.

Observou-se que os resultados de extração dos termos apresentam uma precisão entre 40% e 70%. De acordo com a literatura (12, 13), esses valores são satisfatórios e, portanto, os termos extraídos são adequados para o objetivo de identificação de conceitos na construção automática de hierarquias.

Em resumo, é possível afirmar que uma abordagem linguística como a de busca por sintagmas nominais feita possibilita uma organização do conhecimento em forma de hierarquias de conceitos automaticamente.

É importante salientar que apesar de todos os experimentos neste artigo terem sido realizados sobre o *corpus* de Pediatria, o mesmo processo pode ser feito para qualquer outro *corpus* em português anotado pelo *parser* PALAVRAS. Portanto, um trabalho futuro natural ao descrito neste artigo é o estudo destas abordagens a outros *corpora.*

Outra sequência natural é continuar o processo de construção de ontologias utilizando as listas de termos extraídas e as hierarquias de conceito como base.

Este artigo descreve um processo refinado de extração baseado em informações linguísticas que possibilita a extração e organização do conhecimento de um domínio específico feito por intermédio da construção de hierarquias. Neste sentido, a principal contribuição deste artigo é o uso das informações sintáticas, em especial o uso do núcleo dos sintagmas nominais, para determinar a hierarquia entre os termos automaticamente extraídos.

Este importante passo diminui o caminho a ser percorrido para o objetivo final que é a construção automática de ontologias de um domínio específico. Logo, é melhor ter a hierarquia de conceitos relevantes como ponto de partida, ao invés de construir ontologias a partir de uma grande quantidade de termos que não possuem necessariamente relevância terminológica.

Adicionalmente, uma contribuição imediata deste trabalho é a geração de uma hierarquia sobre o domínio de Pediatria, bem como sua visualização através de árvores hiperbólicas com a ferramenta Treebolic. Esta contribuição pode desde já auxiliar profissionais que atuem em Pediatria a encontrar facilmente relacionamentos entre diversos termos específicos e relevantes desta área.

Referências

1. Buitelaar P, Cimiano P, Magninni B. Ontology learning from text: Methods Evaluation and Applications. IOS Press 2005.
2. Lopes L, Oliveira LHM, Vieira R. Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches Proc. of 9th PROPOR 2010.
3. Bick E. The parsing System “Palavras”. PhD thesis Arhus University 2000.
4. Lopes L, Fernandes P, Vieira R, Fedrizzi G. ExATOlp An Automatic Tool for Term Extraction from Portuguese Language Corpora. Proc. of 4th. LTC’09, p. 427-431, 2009.
5. Lopes L, Vieira R, Finatto MJ, Zanette A, Martins D, Ribeiro Jr LC. 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.
6. Coulthard RJ. The application of Corpus Methodology to Translation. Dissertação de Mestrado, UFSC, 2005.
7. Kuramoto H. Nominal Groups: a New Purpose to Information Retrieval. DataGramaZero - Revista de Ciência da Informação, v.3, n.1, 2002.
8. Hearst M. Automatic acquisition of hyponyms from large text corpora. Proc. of 14th COLING, p.539-545, 1992
9. Harris Z. Mathematical Structures of Language. Wiley, 1968
10. Sanserson M, Croft B. Deriving concept hierarchies from text. Proc. of SIGIR, p. 206-213, 1999
11. Bou B. Treebolic a java applet for hyperbolic rendering of hierarchical data. http://treebolic.sortilege.net/en/index.html.
12. Baptista J, Batista F, Mamede N. Building a dictionary of anthroponyms. Proc. of 7th PROPOR 2006.
13. Hulth A. Enhancing linguistically oriented automatic keyword extraction. HLT-NAACL ACL 2004.