Artigo Original

JHI JOURNAL OF HEALTH INFORMATICS

Poronto: ferramenta para construção semiautomática de ontologias em português

Poronto: tool for semi-automatic ontology construction in portuguese

Poronto: herramienta para construcción semiautomática de ontologías en portugués

Faruk Mustafa Zahra[1], Deborah Ribeiro Carvalho[2], Andreia Malucelli[3]

Descritores:
Inteligência Artificial; Vocabulário Controlado; Computação em Informática Médica

RESUMO

Objetivo: Apresentar uma ferramenta semiautomática para construção de ontologias a partir de textos em português na área da saúde. **Método**: Pesquisa aplicada com abordagem quantitativa, operacionalizada em seis etapas: identificação de ferramentas para aprendizagem de ontologia, identificação de ferramentas de anotação linguística, elaboração do protótipo, avaliação do protótipo, elaboração da versão final da ferramenta e avaliação dos resultados. **Resultados**: Foram realizados três experimentos em domínios diferentes. Os termos extraídos foram avaliados por especialistas nas respectivas áreas, sendo a ferramenta considerada relevante para auxiliar no processo de construção de ontologias. **Conclusão**: Foi comprovada a dificuldade em se construir ontologias semiautomaticamente devido à complexidade envolvida no processo de extração de termos, sendo muito importante a participação do especialista no pós-processamento. Dado que a avaliação do especialista é subjetiva, é preciso selecionar especialistas com critérios padronizados e em quantidade significativa para se ter uma avaliação menos sujeita a falhas.

Keywords: Artificial Intelligence; Vocabulary Controlled; Medical Informatics Computing

ABSTRACT

Objective: Present a tool for semi-automatic ontology construction from texts in Portuguese in the health area. **Method**: Applied research with a quantitative approach, operationalized in six steps: identification of tools for ontology learning, identification of linguistic annotation tools, prototype development, prototype evaluation, preparation of the final version of the tool, and results evaluation. **Results**: Three experiments were conducted in different domains. The extracted terms were evaluated by experts in their respective areas. The tool was considered relevant to support the ontology construction process. **Conclusion**: It was demonstrated the difficulty in constructing ontologies semi-automatically due to the complexity involved in the terms extraction process, being very important the participation of an expert in the post-processing. As the expert's evaluation is subjective, it is necessary to select experts with standardized criteria and in a significant number to have an evaluation with fewer failures.

Descriptores:
Inteligencia Artificial; Vocabulario Controlado; Computación en Informática Médica

RESUMEN

Objetivo: Presentar una herramienta para construcción semiautomática de ontologías a partir de textos en portugués en el área de la salud. **Método**: Investigación aplicada con abordaje cuantitativo, operado en seis etapas: identificación de herramientas para aprendizaje de ontología, identificación de herramientas de anotación lingüística, elaboración del prototipo, evaluación del prototipo, elaboración de la versión final de la herramienta y evaluación de los resultados. **Resultados**: Fueron realizados tres experimentos en dominios diferentes. Los términos extraídos fueron evaluados por especialistas en las respectivas áreas, siendo la herramienta considerada relevante para auxiliar en el proceso de construcción de ontologías. **Conclusión**: Fue comprobada la dificultad en construir ontologías automáticamente debido a la complejidad involucrada en el proceso de extracción de términos, siendo muy importante la participación del especialista en el pos-procesamiento. Dado que la evaluación del especialista es subjetiva, es preciso seleccionar especialistas con criterios estandarizados y en cantidad significativa para tener una evaluación menos sujeta a fallas.

[1] Mestre do Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná - PUCPR, Curitiba (PR), Brasil.
[2] Doutora. Docente do Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná - PUCPR, Curitiba (PR), Brasil.
[3] Doutora. Docente do Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná - PUCPR, Curitiba (PR), Brasil.

Autor Coorespondente:**Faruk Mustafa Zahra**
e-mail: **farukz@gmail.com**

Artigo recebido: **14/11/2012**
Aprovado: **05/03/2013**

INTRODUÇÃO

Processar informações usando ontologias, que proveem excelente contexto para o entendimento das informações, tanto para usuários humanos quanto para agentes de software, vem se tornando uma tendência em várias áreas e tipos de aplicações. Entre as áreas mais comuns de aplicação de ontologia estão a medicina e a biologia. A principal razão para isso pode ser a complexidade do conhecimento médico e biológico tornar difícil a confecção de sistemas tradicionais, pois para assistir tarefas médicas, os sistemas precisam de muito conhecimento e capacidade de inferência[1-2].

Há muito interesse no desenvolvimento de ontologias na área da saúde, porém, há pouco trabalho desenvolvido e pouca utilização de ontologias em larga escala. Algumas das razões para isso são o tempo e custo associados neste desenvolvimento. Para a criação de uma ontologia é necessário que um especialista da área em questão transfira o respectivo conhecimento consensual para a ontologia, porém trata-se de um processo extremamente custoso. Além disso, as ontologias devem ser compartilhadas por um grupo de pessoas ou por uma comunidade, o que amplia a dificuldade para a construção por envolver diferentes pessoas com pontos de vistas, muitas vezes, divergentes[3].

Este fato levou ao surgimento de uma nova área de pesquisa denominada de aprendizagem de ontologias (ontology learning), que é definida como um conjunto de métodos e técnicas para a construção semiautomática de uma nova ontologia ou para o enriquecimento de uma já existente.

Porém, para construir uma ontologia de maneira semiautomática é necessário a automatização do processo de aquisição de conhecimento e, para isso, diversas abordagens foram sugeridas[4]: construção de ontologias a partir de texto; de dicionários; de base de conhecimento; de esquemas semiestruturados; e de bases de dados relacionais.

Considerando a grande quantidade de documentos digitais disponíveis na Web e os que cada profissional e organizações possuem, o melhor seria considerá-los como um recurso válido para aquisição de conhecimento para a construção de ontologias. Segundo Buitelaar et al.[5] a utilização de textos como fonte de aquisição de conhecimento parece ser um caminho correto, visto que a linguagem é a primeira forma de transferência de conhecimento entre os seres humanos. Porém, a precisão e consistência não podem ser garantidas com métodos automatizados, sendo necessário sempre o pós-processamento por seres humanos.

Face ao exposto, o presente artigo tem como objetivo apresentar uma ferramenta para construção semiautomática de ontologias na área da saúde a partir de textos em português.

MÉTODO

Pesquisa aplicada com abordagem quantitativa, operacionalizada em seis etapas, de acordo com a Figura 1:

Na etapa 1, identificação de ferramentas para aprendizagem de ontologia, foi realizada uma busca na Web pelas ferramentas para aprendizagem de ontologia, com o objetivo de avaliá-las para identificar qual era a mais adequada para ser utilizada como base para este trabalho. Das dezoito ferramentas localizadas, apenas três estavam disponíveis para download e instalação. Foram avaliadas as ferramentas Keyphrases Extraction Algorithm (KEA)[6], TERMINAE[7] e Text-To-Onto[8], sendo selecionada a Text-to-Onto. Na etapa 2, identificação de ferramentas de anotação linguística, foram avaliadas as funcionalidades disponíveis em softwares existentes no portal da Linguateca[9]. Foram avaliadas as ferramentas TreeTagger[10] e VISL[11], sendo selecionada o TreeTagger por ser gratuita. O protótipo, na etapa 3, foi desenvolvido para plataforma desktop utilizando o padrão swing da linguagem Java, versão 6, por meio do IDE (Integrated Development Environment) Eclipse. Para a avaliação do protótipo, etapa 4, foram utilizados como corpus 12 artigos científicos da área de Histocompatibilidade, sendo avaliado a facilidade de uso, facilidade de distribuição e qualidade dos termos extraídos automaticamente.

A versão final da ferramenta, etapa 5, foi desenvolvida em tecnologia de código aberto e gratuito, Java versão 6,

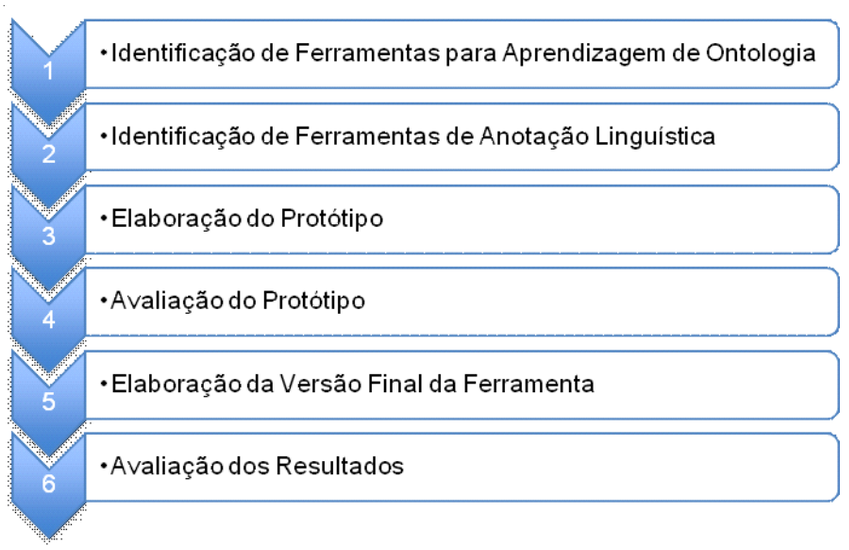

Figura 1 – Etapas do Método

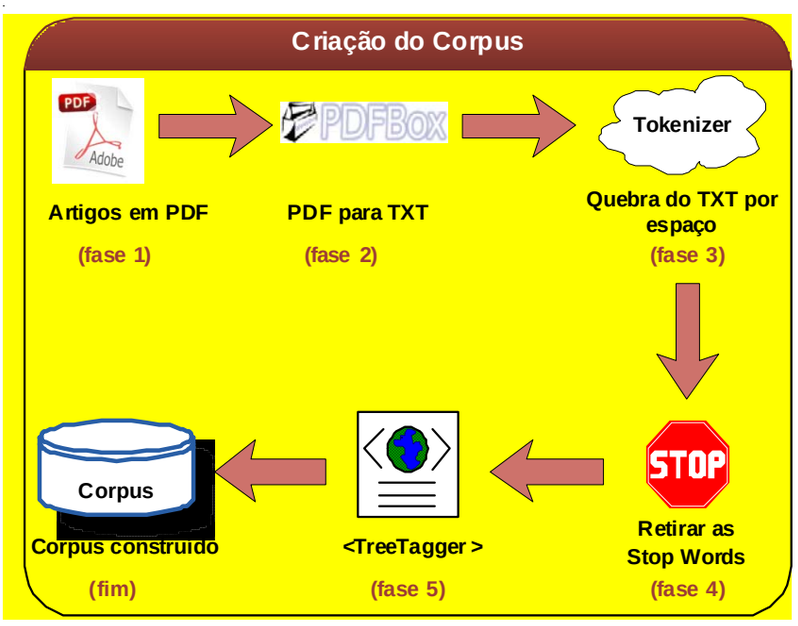

Figura 2 - Processo da criação do corpus da ferramenta

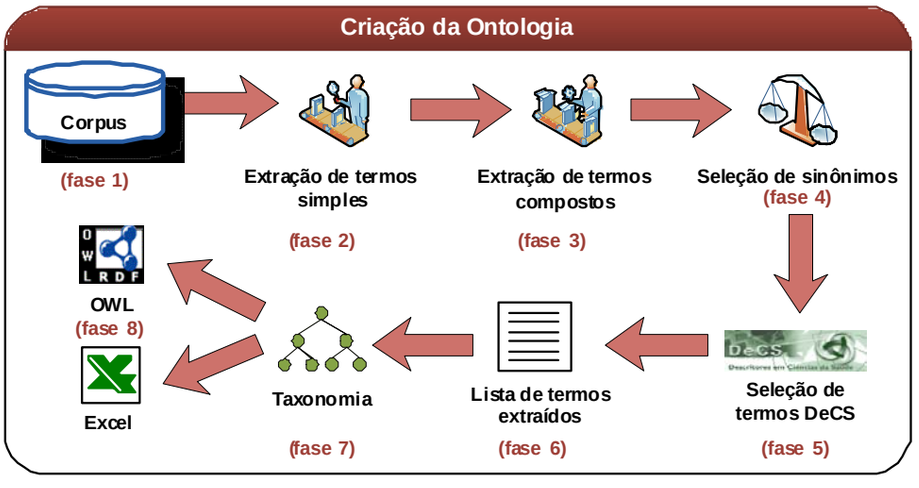

Figura 3 - Processo da criação da ontologia

Java Server Faces (JSF), Apache MyFaces[12] e Jboss RichFaces[13]. O processo de geração semiautomática de ontologias foi dividido em duas etapas: a criação do corpus e a criação da ontologia.

A criação do corpus é dividida em cinco fases, de acordo com a Figura 2: (fase 1) o usuário envia os artigos que deseja processar em formato PDF; (fase 2) os artigos são transformados em arquivos texto limpo (sem os marcadores padrões de arquivos PDF), utilizando o PDFBox[14]; (fase 3) ocorre um pré-processamento dos textos, onde o texto é dividido por espaços em branco para posteriormente ser feito o processamento de anotação linguística com o TreeTagger[10]; (fase 4) as Stop Words são removidas; (fase 5) os textos são processados com o TreeTagger.

Com o corpus construído, a etapa de criação da ontologia pode ser iniciada, a qual é dividida em oito fases, de acordo com a Figura 3:

Na fase 1, o usuário preenche os filtros para processar o corpus com a quantidade mínima de vezes em que um termo simples aparece no corpus, número mínimo de termos compostos, número mínimo de vezes em que um termo composto aparece no corpus; seleção dos termos compostos, inclusão ou não no resultado dos termos compostos; se apenas substantivos, inclusão no resultado apenas de termos marcados pelo TreeTagger como substantivos ou se todos os termos; utilização ou não da medida tf-idf[15] como medida de seleção; utilização ou não da medida de entropia[16] , como medida de seleção. De acordo com Wiener, 1948 "a soma de informação em um sistema é a medida de seu grau de organização; a entropia é a medida de seu grau de desorganização; um é o negativo do outro". Os termos simples são extraídos na fase (2) aplicando as medidas frequência[15], tf-idf[15] e entropia[16]. Na fase (3) os termos compostos são extraídos com base em regras expressas por sequências de tipos morfológicos de acordo com o Quadro 1 e são aplicadas as medidas frequência, tfi-df e entropia.

Na fase 4 uma busca por sinônimos dos termos é

Quadro 1 - Regras de identificação de termos compostos adotadas na ferramenta

Regras de sequência morfológica
SUbstantivo + **Ad**jetivo
SUbstantivo + **PR**eposição + **Su**bstantivo
SUbstantivo + **PR**eposição + **AD**jetivo + **SU**bstantivo
SUbstantivo + **PR**eposição + **SU**bstantivo + **PR**eposição + **SU**bstantivo

Resultados

Lemma	Total	Tfidf	Entropy	Decs	Sinônimos
câncer	746	214.62	1.11	✓	câncer...
caso	485	139.53	1.04	✗	acontecimento...
taxa	386	111.05	1.05	✗	contribuição...
bruto	313	216.96	1.01	✗	atroz...
estimativa	291	83.72	1.03	✗	
localização	285	81.99	1.03	✓	
incidência	270	77.68	1.07	✓	
novo	223	154.58	1.02	✗	actual...
saúde	220	63.3	1.08	✓	
brasil	217	62.43	1.12	✓	
neoplasia	215	61.86	1.05	✗	
pele	205	58.98	1.06	✓	couro...
primária	200	138.63	1.01	✗	
mulher	177	50.92	1.08	✗	dama...
tabela	168	116.45	1.03	✗	tabela...
melanoma	146	202.4	1.0	✓	
homem	141	40.57	1.07	✗	homem...
pulmão	130	37.4	1.08	✓	
risco	123	85.26	1.07	✓	linha...
população	105	30.21	1.11	✓	

« « 1 2 3 4 5 6 7 8 9 10 » »»

Figura 4 - Lista dos termos para seleção

realizada na lista do OpenThesaurusPT[17] para facilitar o critério de seleção do termo pelo usuário. Uma pesquisa é realizada na fase 5 para verificar se os termos extraídos possuem correspondência na lista de Descritores em Ciência da Saúde (DeCS), também para facilitar o critério de seleção do termo pelo usuário. Na fase 6 os termos simples e os compostos, extraídos pela ferramenta, são apresentados ao usuário, assim como alguns sinônimos para estes termos e se o termo está ou não incluído na lista de descritores. A partir da apresentação (Figura 4) devem ser selecionados pelo usuário os termos mais relevantes a ser inseridos na ontologia.

Após selecionar os termos relevantes, o usuário pode optar pela organização destes termos em uma taxonomia, fase 7, que é construída com um método baseado em termos compostos. O usuário pode exportar o resultado para o formato XLS ou OWL, fase 8, selecionando a opção do Menu "Exportar para...". A Figura 5 apresenta um exemplo de parte da ontologia gerada, no editor de ontologias Protégé. Com isso o processo de construção da ontologia é finalizado.

A avaliação dos resultados, etapa 6 do método, foi realizada a partir de textos em português de três áreas da saúde: câncer de mama, sendo processados 8 textos do Ministério da Saúde e 16 textos do Instituto Nacional do Câncer (INCA), totalizando 24 arquivos no corpus; eventos adversos pós-vacinação, sendo processado o Manual de Vigilância Epidemiológica de Eventos Adversos Pós Vacinação[18]; e inventário vocabular de enfermagem, sendo processado um texto referente ao cenário da força de trabalho de Agentes de Saúde, cujo objetivo original com este texto era a contribuição para uma norma mundial de enfermagem em saúde coletiva[19]. A execução dos experimentos iniciou com o processamento do corpus pela ferramenta, sendo contabilizada a quantidade de termos repetidos, ou seja, termos que se repetiram mais

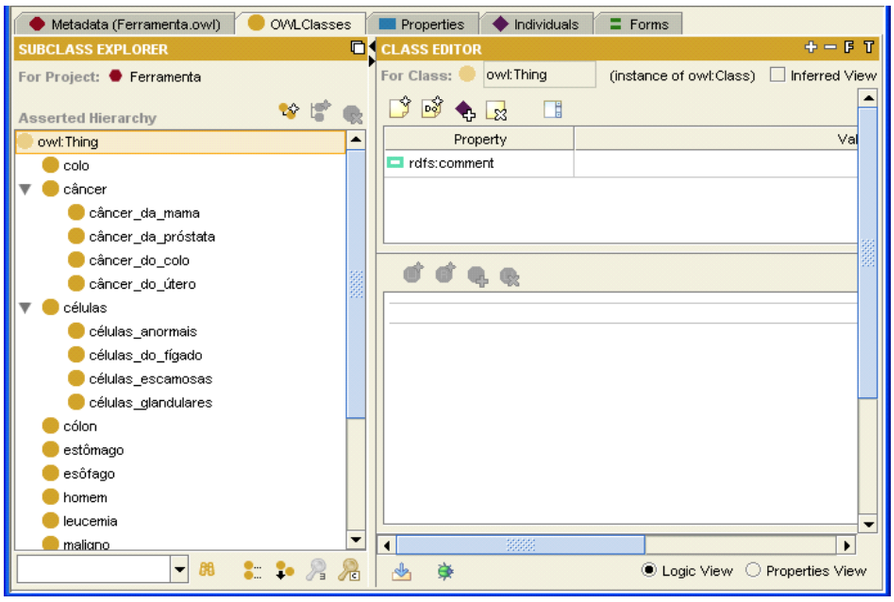

Figura 5 - Ontologia importada para o Protégé

de uma vez no corpus; a quantidade de termos únicos, ou seja, sem levar em consideração a quantidade de vezes em que aparecem no corpus; e a quantidade dos termos do corpus extraídos pela ferramenta. O processo para avaliar se os termos extraídos correspondiam aos termos relevantes na área foi subjetivo, sendo consultados três especialistas de cada uma das áreas. Cada especialista recebeu uma planilha eletrônica com os termos extraídos automaticamente pela ferramenta e foi solicitado que eles assinalassem os termos que eles consideravam relevantes, de acordo com o seu conhecimento. Foram considerados validados os termos que obtiveram uma porcentagem de 66% de concordância, ou seja, onde dois dos três especialistas concordaram com o termo.

RESULTADOS E DISCUSSÃO

A Figura 6 apresenta a página principal da ferramenta, com a qual foram realizados experimentos com textos nas áreas de câncer de mama (experimento 1), eventos adversos pós-vacinação (experimento 2) e inventário vocabular (experimento 3).

No experimento 1 haviam 481.683 termos repetidos no corpus, sendo destes 55.207 termos únicos, dos quais o PORONTO selecionou 1.442 termos. No experimento 2 haviam 35.465 termos repetidos no corpus, sendo destes 6.169 termos únicos e 1.090 extraídos pela ferramenta; e o experimento 3 tinha 3.309 termos repetidos, sendo 681 termos únicos e a ferramenta extraiu 428 termos.

Estes resultados se devem à poda realizada pelos filtros que eliminaram termos simples e compostos com frequência menor que o parâmetro inserido no filtro e termos simples que não eram substantivos.

Foram selecionados pelos especialistas como termos válidos 449 termos no experimento 1; 370 termos no experimento 2; e 190 termos no experimento 3, ou seja, a ferramenta extraiu corretamente 31,14%; 33,94%; e 44,39% em cada experimento, respectivamente.

O Quadro 2 apresenta a distribuição dos termos selecionados, nos três experimentos, levando em consideração a medida de frequência.

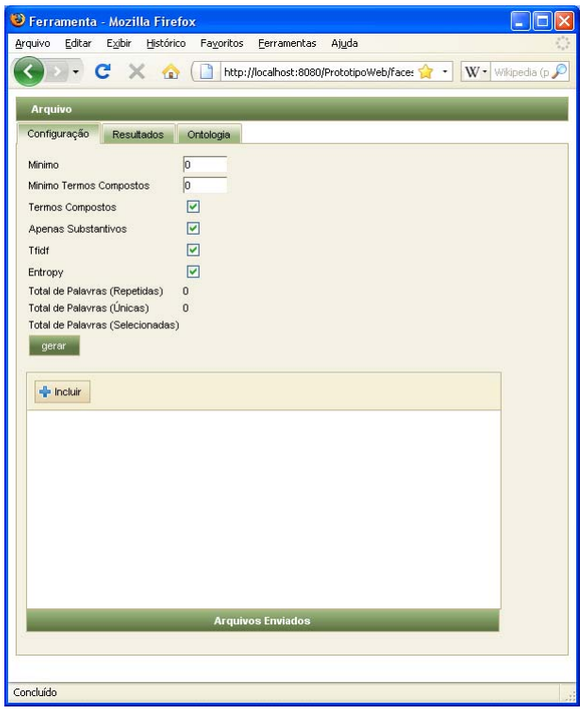

Figura 6 - Página de entrada

Pode-se concluir que mais da metade dos termos selecionados pelos especialistas aparecem com menor frequência no corpus. Isso sugere a criação de um limiar de corte com o número máximo de vezes que um determinado termo pode aparecer em um corpus, além do número mínimo de vezes adotado neste trabalho.

A Figura 7 apresenta a quantidade de termos compostos extraídos e selecionados em cada experimento. Após a contabilização dos termos compostos extraídos e selecionados, pode-se contabilizar também a quantidade de termos compostos extraídos e selecionados de acordo com cada regra.

Quadro 2 - Frequência dos termos selecionados

	Câncer de mama		Eventos adversos pós-vacinação		Inventário vocabular de enfermagem	
Frequência dos termos selecionados	30 - 100 vezes	289 (64,37%)	3 - 50 vezes	342 (92,43%)	1 - 5 vezes	160 (84,21%)
	100 - 1000 vezes	148 (32,96%)	50 - 100 vezes	13 (3,51%)	5 - 10 vezes	23 (12,11%)
	mais de 1000 vezes	12 (2,67%)	mais de 100 vezes	15 (4,05%)	mais de 10 vezes	7 (3,68%)

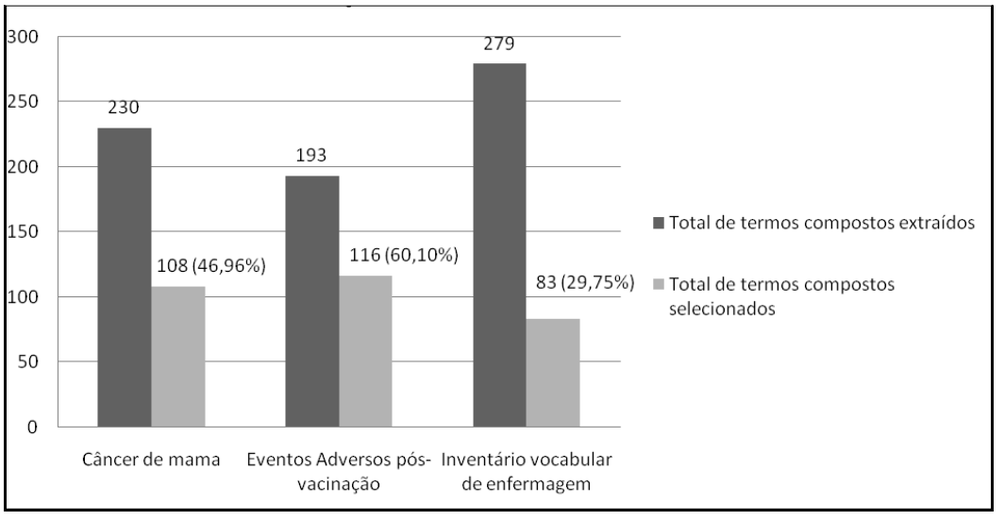

Figura 7 - Termos compostos extraídos e selecionados

Os Quadros 3, 4 e 5 apresentam a porcentagem de acerto no experimento 1, experimento 2 e experimento 3, respectivamente, sendo SU = SUbstantivo, AD = ADje

A partir da análise dos resultados pode-se concluir que a quantidade de termos extraídos pela ferramenta é dependente da poda realizada pelos filtros que eliminam termos simples e compostos de acordo com a frequência estabelecida. A alta frequência com que um termo aparece no corpus não significa que ele será selecionado.

Quadro 3 - Porcentagem de acerto em cada regra de extração de termos compostos no experimento 1

Câncer de mama			
Regra	**Extraídos**	**Selecionados**	**Total de acertos**
SU AD	95	48	50,53%
SU PR SU	135	60	44,44%
SU PR AD SU	0	0	0,00
SU PR SU PR SU	0	0	0,00
Total	230	108	-

Quadro 4 - Porcentagem de acerto em cada regra de extração de termos compostos no experimento 2

Eventos Adversos pós-vacinação			
Regra	**Extraídos**	**Selecionados**	**Total de acertos**
SU AD	153	106	69,28%
SU PR SU	40	10	25,00%
SU PR AD SU	0	0	0,00
SU PR SU PR SU	0	0	0,00
Total	193	116	-

Muitos dos termos extraídos de maneira automática pela ferramenta não foram selecionados como relevantes pelos especialistas devido à quantidade de "ruído" no corpus. Sendo assim, é importante incluir outros tipos de filtros, como por exemplo, para exclusão de nomes próprios do corpus, sendo possível assim refinar ainda mais a extração dos termos.

Quadro 5 - Porcentagem de acerto em cada regra de extração de termos compostos no experimento 3

Inventário Vocabular de Enfermagem			
Regra	**Extraídos**	**Selecionados**	**Total de acertos**
SU AD	92	28	30,43%
SU PR SU	147	46	31,29%
SU PR AD SU	0	0	0,00
SU PR SU PR SU	40	9	22,50%
Total	279	83	-

Após a seleção dos termos pelos especialistas foi possível contabilizar também a quantidade destes termos que são DeCS, Quadro 6. Nos três experimentos mais de 20% dos termos selecionados são DeCS, sendo possível concluir que nos três experimentos os descritores tiveram um peso fundamental para a extração dos termos automaticamente pela ferramenta, podendo ser utilizada como medida de extração. No experimento 1, dos 449 termos selecionados pelos especialistas, 110 são DeCS; no experimento 2, dos 370, 144 termos são DeCS; e no experimento 3, dos 190, 49 termos são DeCS.

Notou-se durante a avaliação dos experimentos uma diferença entre a quantidade de termos selecionados pelos especialistas, destacando a subjetividade do processo. É possível que um determinado especialista conheça mais sobre o domínio, tendo assim mais confiança na seleção dos termos, ou ser mais rigoroso quanto ao critério de seleção. Não é escopo deste trabalho analisar os critérios de seleção de termos, mas salientar a discussão sobre a subjetividade envolvida dentro do mesmo domínio e escopo.

Quadro 6 - Quantidade dos termos selecionados que são DeCS

	Câncer de mama	Eventos adversos pós-vacinação	Inventário vocabular de enfermagem
Total de termos selecionados pelos especialistas	449 (31,14%)	370 (33,94%)	190 (44,39%)
Total de termos selecionados pelos especialistas que são DECS	110 (24,50%)	144 (38,92%)	49 (25,79%)

Com isso concluiu-se, como destacado por Dellschaft e Staab[20] , que os fatores que mais influenciam a avaliação são: a escolha correta dos especialistas e a quantidade deles, pois escolhendo com critério os especialistas tem-se a certeza de que os termos serão selecionados corretamente e com uma quantidade maior de especialistas eliminam-se as possíveis falhas geradas por eles.

CONCLUSÃO

Este trabalho apresentou uma ferramenta para construção semiautomática de ontologias a partir de textos em português na área da saúde, com algumas funcionalidades que a diferenciam das já existentes, dado que não há necessidade do usuário realizar previamente a anotação linguística do corpus, estando este acoplado à ferramenta, facilitando sua usabilidade; a ferramenta está disponível em um ambiente Web, onde o acesso é facilitado para profissionais e para a comunidade cientifica; há inserção de sinônimos, facilitando e auxiliando na seleção dos termos; e há inserção de termos relacionados com os DeCS, facilitando também a seleção dos termos pelo especialista.

Analisando-se os resultados dos experimentos conclui-se que é possível construir ontologias na área da saúde, de maneira semiautomática, a partir de textos em português, tendo como principais contribuições: ferramenta para a construção semiautomática de ontologias a partir de textos em português; avaliação prática de ferramentas disponíveis para construção semiautomática de ontologias; avaliação prática de ferramentas disponíveis para anotação linguística; construção de três novas estruturas ontológicas na área da saúde.

Na aquisição de conhecimentos, os engenheiros ontológicos possuem dificuldades em capturar e representar o conhecimento a ser representado na ontologia, sendo necessário um especialista do domínio para transferir este conhecimento, para que posteriormente possa ser formalizado e implementado em uma linguagem interpretável por máquina. Este é um processo extremamente moroso e custoso, sendo que a construção das três estruturas ontológicas exemplifica a aplicação prática da ferramenta, que permite que o conhecimento seja extraído mais rapidamente, deixando de ser um totalmente manual.

Foi comprovada a dificuldade envolvida na construção semiautomática de ontologias devido à complexidade inerente ao processo de extração de termos, sendo fundamental a participação do especialista na etapa do pós-processamento. É possível também concluir que, tendo em vista a avaliação do especialista ser subjetiva, é preciso selecionar especialistas com critérios bem definidos e padronizados e em quantidade significativa para se ter uma avaliação menos sujeita a falhas, oportunizando uma ontologia de consenso.

A partir deste estudo, surgem perspectivas de trabalhos futuros, entre as quais: melhorar a poda do corpus retirando: nomes próprios, referências, quadros, tabelas e outros que possam gerar ruído no processamento; utilizar o DeCS como medida de extração; criar um filtro com limiar de corte com o máximo de vezes em que um termo pode aparecer em um corpus; implementar novas regras de identificação de termos compostos; implementar novos métodos para geração taxonômica; e implementar a extração de relacionamentos (não hierárquicos).

REFERÊNCIAS

1. Musen MA. Medical informatics: searching for underlying components. Methods Inf Med. 2002;41(1):12-9.
2. Freitas F, Schulz S. Ontologias, web semântica e saúde. Rev Electron Comun Inf Inov Saude [online]. 2009:3(1) [acesso em 2009 Nov 5]. Disponível em: http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/238/246
3. Cimiano P. Ontology Learning and Population: Algorithms, evaluation and applications [thesis]. Karlsruhe (Germany): University of Karlsruhe; 2005.
4. Maedche A, Staab S. Ontology learning for the semantic web. IEEE Intell Syst. 2002;16(2):72-9.
5. Buitelaar P, Cimiano P, Magnini B. Ontology learning from text: an overview. In: Buitelaar P, Cimiano P, Magnini B. (Ed.). Ontology learning from text: methods, evaluation and applications. Amsterdan: IOS Press; 2005. p.3-11.
6. Jones S, Paynter GW. Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. J Am Soc Inf Sci Technol. 2002;53(8):653-77.
7. Biébow B, Szulman S. TERMINAE: a linguistic-based tool for the building of a domain ontology. In: EKAW'99 Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and management. Dagstuhl, Germany, LCNS, Berlin; 1999. p.49-66.
8. Maedche A, Staab S. Ontology learning. In: Staab S, Studer R (Eds.). Handbook on ontologies in information systems. 2003 [acesso em 2009 nov 3]. Disponível em: http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/handbook-ontology-learning.pdf
9. Linguateca. Um centro de recursos distribuído para o processamento computacional da língua portuguesa. [acesso em 2009 set 01]. Disponível em: http://www.linguateca.pt
10. Schmid H. Tree tagger - a language independent part-of-speech tagger. 1994 [acesso em 2009 nov 03]. Disponível em: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
11. VISL - Visual Interactive Syntax Learning. [acesso em 2009 out 15]. Disponível em: http://visl.hum.sdu.dk/visl/pt
12. Myfaces. Apache Myfaces - project of the Apache Software Foundation. [acesso em 2009 set 01]. Disponível em: http://myfaces.apache.org
13. Richfaces. JBoss Richfaces component library. [acesso em

2009 set 01]. Disponível em: http://www.jboss.org/jbossrichfaces
14. PDFBOX. Apache PDFBox - Java PDF Library. [acesso em 2009 out 15]. Disponível em: http://incubator.apache.org/pdfbox
15. Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge, Massachusetts: The MIT Press; 1999.
16. Wiener N. Cybernetics. In: Foerster HV. Cybernetics of cybernetics: the control of control and the communication of communication. Minneapolis: Future Systems; 1948. p. 7-17.
17. OpenThesaurusPT. Um projeto Open Source para a construção de um Dicionário de Sinônimos para a língua portuguesa. [acesso em 2009 out 15]. Disponível em http://openthesaurus.caixamagica.pt
18. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. Manual de vigilância epidemiológica de eventos adversos pós-vacinação. 2ª ed. Brasília: Ministério da Saúde; 2008.
19. Garcia TR, Nóbrega MML da. Inventário vocabular de fenômenos e ações de enfermagem. In: Garcia TR, Nóbrega MML da. (Org.). Sistemas de classificação em Enfermagem: um trabalho coletivo. João Pessoa(PB): Idéias; 2000. p. 83-170.
20. Dellschaft K, Staab S. Strategies for the evaluation of ontology learning. Amsterdan: IOS Press; 2007.