Extração automática de termos candidatos às ontologias: um estudo de caso no domínio da hemoterapia

Fabrício M. Mendonça[1], Maurício B. Almeida[1],
Renato R. Souza[2], Daniela L. Silva[1,3]

[1]Escola de Ciência da Informação – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 – Campus Pampulha – 31.270-901 – Belo Horizonte – Brasil

[2]Escola de Matemática Aplicada – Fundação Getúlio Vargas
Praia do Botafogo – CEP – Rio de Janeiro – Brasil

[3]Departamento de Biblioteconomia – Universidade Federal do Espírito Santo
Av. Fernando Ferrari, 514 - Goiabeiras – 29.075-910 – Vitória – Brasil

fabriciommendonca@gmail.com, mba@eci.ufmg.br, renato.souza@fgv.br, danielalucas@hotmail.com

Abstract. *This paper describes a case study conducted within the domain of blood transfusion aiming at non-exhaustively extraction of candidate terms for an ontology of human blood. The process involved both the construction of a corpus and its automatic processing, and the retrieval of specialized terms. As our main result, we have obtained candidate medical terms to be used in a ontology of blood transfusion processes.*

Resumo. *O presente artigo descreve um estudo de caso realizado no domínio de hemoterapia para a extração automática e não exaustiva de termos candidatos às ontologias sobre o sangue humano. O processo envolveu a construção de um corpus, seu processamento por ferramenta automática e a recuperação de termos médicos. Ao final do experimento, obtiveram-se os termos médicos candidatos a ontologia sobre processos hemoterapêuticos.*

1. Introdução

A ampla utilização da internet e das novas tecnologias de informação, tais como os dispositivos móveis, as redes sociais, os sistemas de informação eletrônicos, têm mudado a forma do ser humano manipular informação. Uma das consequências destas mudanças é a necessidade de novas abordagens para organização da informação.

No âmbito da medicina, a extração de informação a partir de textos tem sido abordada de duas formas principais [Friedman e Hripcsak 1999]: (i) a abordagem manual – por vezes denominada curadoria – realizada por especialistas capacitados em abstrair informações dos textos médicos e correspondê-los a conceitos de terminologias médicas; e (ii) a abordagem automática, que consiste na extração de termos médicos realizada, normalmente por ferramentas de processamento de linguagem natural (PLN).

Considerando as possíveis potencialidades da extração automática na organização e representação da informação médica [Winnenburg et al. 2008], o presente

artigo descreve um estudo de caso no domínio da hemoterapia, em que foi realizada a extração de termos médicos de um corpus para uso em ontologia sobre os processos da manipulação do sangue humano [Almeida et al. 2010]. Optou-se por utilizar ontologias para modelagem do conhecimento na área, devido às suas vantagens significativas em relação à modelagem conceitual tradicional, caracterizada por ser *ad hoc* e orientada por casos [Smith e Welty 2001].

A ferramenta *Sketch Engine*[1] foi utilizada para a construção e o processamento morfossintático de um corpus sobre o sangue humano, a partir de textos especializados. A extração de termos se baseou em uma lista de sufixos médicos relevantes no domínio, sugerida por especialistas com base em seus conhecimentos, além de referências da área.

O presente artigo está estruturado nas seguintes seções: na seção 2 discorre-se sobre o uso de ferramentas de PLN para extração de informação e construção de ontologias; na seção 3 apresenta-se a metodologia usada para a extração de termos do corpus como possíveis candidatos à ontologia; na seção 4 são descritos os resultados obtidos com o experimento realizado; e, na seção 5 são apresentadas conclusões acerca deste trabalho e os trabalhos futuros previstos nesta pesquisa.

2. Uso de ferramentas de PLN na extração e organização da informação

Embora existam ferramentas de PLN disponíveis para extração de informação e auxílio na construção de ontologias [Buitelaar et al. 2003], [Cimiano e Volker 2005], [Wätcher e Schroeder, 2010], os resultados nem sempre são satisfatórios para tratar a complexidade envolvida na aquisição de conhecimento para ontologias. Na maioria dos casos, torna-se necessária a intervenção humana para ajuste dos termos à ontologia [Smith et al. 2005]. Ainda assim, as ferramentas de PLN são consideradas úteis para diversas tarefas na construção de ontologias e no processo de curadoria [Buitelaar et al 2003] [Winnenburg et al 2008], principalmente na fase de aquisição de conhecimento.

Afora essa discussão, fundamental é ressaltar aspectos essenciais para a extração de informação de textos através das ferramentas de PLN, que se referem: à definição ou criação de um corpus no domínio sob estudo e aos procedimentos usados para a análise e anotação linguística do corpus (processamento). Um corpus é "um conjunto estruturado de grandes dimensões de textos, eletronicamente armazenados e processados, usado para um propósito definido e que seja representativo do domínio sob estudo" [McEnery e Wilson 2011]. Nesse sentido, nem todo conjunto de textos eletrônicos é propriamente um corpus. De fato, os critérios relevantes para a construção de corpora envolvem *autenticidade, tamanho, amostragem, representatividade* e *balanceamento* [Biber 1993] [Tognini-Bonelli 2001].

Outro aspecto fundamental refere-se ao trabalho de análise linguística do corpus, que assim como sua criação, pode ser feito manualmente ou com o uso de ferramentas automatizadas. Essa análise envolve imprescindivelmente a etapa de codificação do corpus, bem como a anotação ou etiquetagem dos seus elementos. Para o processo de anotação de corpora também existem princípios como a *recuperabilidade* e a *capacidade de extração* [Leech 1993].

[1] A ferramenta está disponível em: http://Sketch Engine.co.uk/. Acesso em 29 de Junho de 2012.

3. Metodologia para extração dos termos candidatos

O objetivo da presente seção é apresentar a metodologia empregada para extrair do corpus formado os termos candidatos à ontologia de processos do sangue humano. A abordagem utilizada é considerada semi-automática, devido à intervenção humana, e envolveu três etapas principais: (i) construção de um corpus no domínio do sangue, utilizando-se uma ferramenta automática (seção 3.1); (ii) processamento automático do corpus através de análise morfossintática (seção 3.2); e (iii) cálculo da frequência dos termos candidatos à ontologia (seção 3.3), por meio de tarefas manuais e automáticas.

3.1. Construção de um corpus no domínio do sangue

Os textos escolhidos para compor o corpus no domínio do sangue foram extraídos de um manual técnico sobre os padrões de qualidade para manipulação do sangue humano da instituição americana AABB[2] *Technical Manual*, 17ª edição. Nesse sentido, procura-se atender aos critérios citados para construção de corpus.

Com a amostra de textos selecionada partiu-se para a criação do *corpus* no *Sketch Engine*, que é capaz de processar textos em formato *pdf* para construção de corpora. Dos 32 capítulos da 17ª edição do *AABB's Technical Manual*, 27 foram incluídos na formação do corpus, totalizando 369.741 *tokens* identificados.

3.2. Análise morfossintática do *Blood Corpus*

A análise morfossintática do corpus foi realizada utilizando-se também a ferramenta *Sketch Engine*. Essa etapa permitiu a identificação e anotação linguística dos *tokens* do *Blood Corpus* (BC).

Para anotação do corpus, o *Sketch Engine* utiliza: (i) a linguagem de marcação XML na anotação das informações linguísticas dos elementos do texto; e (ii) princípios de anotação de corpora do padrão internacional *Text Encoding Initiative*[3] (TEI). Já o tipo de anotação, realizada pela ferramenta no corpus BC, corresponde à anotação *Part-of-Speech (POS) Tagger*, que inclui a etapa de lematização e a anotação das categorias morfo-sintáticas dos elementos do texto. Nessa anotação, cada item lexical é associado a apenas uma categoria gramatical (etiqueta) de acordo com seu uso na frase.

3.3. Cálculo da frequência dos termos candidatos a processos do sangue

Após a construção e a anotação do corpus, o passo seguinte foi extrair termos candidatos à ontologia dos processos sobre o sangue humano, conforme segue.

A estratégia semi-automática utilizada envolveu três passos: (i) seleção manual de sufixos médicos que identifiquem processos; (ii) cálculo automático da frequência dos termos que possuem tais sufixos no corpus; e (iii) agrupamento manual dos termos recuperados em classes semânticas de acordo com o sufixo que possuem.

---

[2] AABB é uma associação internacional que conta com 2000 instituições de saúde e 8000 profissionais vinculados, originários de 80 diferentes países do mundo [AABB 2012].

[3] O *Text Encoding Initiative* (TEI): envolve três importantes associações de linguística computacional do mundo, para a criação de formatos padronizados de anotação [McEnery e Wilson 2001].

A justificativa pela escolha dos sufixos dos termos como base para cálculo de frequência e, consequentemente, para extração do corpus, deve-se ao fato de que essa parte da palavra, normalmente, representa o significado (semântica) de um termo médico. Desta forma, o **sufixo** indica o procedimento, a condição ou a doença representada pelo termo. Em "*polycythaemia*", por exemplo, o prefixo *poly* indica *muitos,* a raiz *cythea* representa *célula* como parte do corpo humano onde o processo ocorre e o sufixo *emia* é relativo *à falta de algo*.

Nesse sentido, consultaram-se especialistas e referências na área para a criação de uma lista de sufixos que representem procedimentos médicos na área de hemoterapia. Os sufixos selecionados foram: *-apheresis, -centesis, -desis, -ectomy, -opsy, -rrhaphy, -metry, -scopy, -oscopy, -otomy, -ostomy, -pexy* e *-plasty*.

De posse dos sufixos médicos, procedeu-se com a construção de expressões regulares utilizando a linguagem *Corpus Query Language* (CQL) dentro da ferramenta *Sketch Engine*. A execução de tais expressões na ferramenta permitiu recuperar todos os termos do corpus que possuem esses sufixos, bem como a frequência absoluta do termo no corpus, ou seja, seu número de ocorrências.

O passo seguinte para a extração dos termos do corpus BC foi selecionar da lista de frequência calculada apenas aqueles termos com maior frequência e agrupá-los em sua classe semântica correspondente, de acordo com o significado do seu sufixo. A execução deste último passo possibilitou a sugestão de termos candidatos à ontologia.

4. Resultados parciais

Nesta seção, apresentam-se os resultados obtidos até o momento com a extração automática de termos do *corpus* e sua representação como classes de uma ontologia sobre os processos de manipulação do sangue humano.

O agrupamento manual dos termos recuperados do corpus BC em classes semânticas, de acordo com os sufixos que os compõem, é mostrado na tabela 1. Nela, os números entre parênteses indicam a frequência do termo no corpus. Considerou-se como frequência mínima para este agrupamento valores maiores ou iguais a três ocorrências, assim os demais termos recuperados foram descartados.

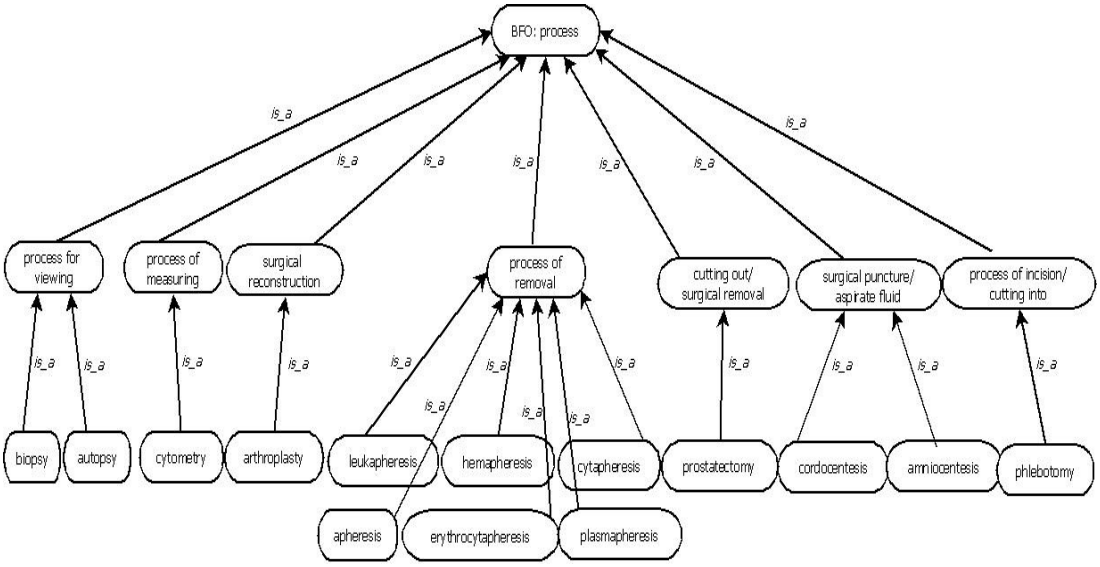
Tabela 1: Agrupamento dos termos recuperados em classes semânticas

Classe semântica	Termo recuperado (frequencia)
Process of removal	*apheresis (124), plasmapheresis (15), leukapheresis (12), hemapheresis (6), Leukapheresis (6), rythrocytapheresis (5), Plasmapheresis (3), Cytapheresis (3)*
Surgical puncture or aspirate fluid	*cordocentesis (16), amniocentesis (14), Amniocentesis (4)*
Process of incision or cutting into	*phlebotomy (32), Phlebotomy (9)*
Process of measuring	*cytometry (20)*
Process for viewing	*biopsy (7), autopsy (5)*
Surgical reconstruction	*arthroplasty (9)*
Cutting out	*prostatectomy (8)*

As classes semânticas e os termos mostrados na tabela 1 correspondem aos candidatos à ontologia sobre os processos do sangue humano. A fim de incluí-los em tal

ontologia, foi construída uma taxonomia desses elementos (vide figura 1), que representam processos hemoterapêuticos.

Figura 1: Taxonomia dos tipos de processos hemoterapêuticos



Para a construção da taxonomia partiu-se da utilização de uma ontologia de fundamentação - a *Basic Formal Ontology* (BFO) [Grenon e Smith 2004] – que representa, normalmente, uma boa prática na construção de ontologias de domínio. Assim a taxonomia inicia-se com a classe *process* da BFO e, na sequência, temos as classes correspondentes aos processos hemoterapêuticos: (i) no segundo nível, são representados os processos mais gerais, agrupados de acordo com o sufixo; e (ii) no terceiro nível, temos os processos específicos, que correspondem exatamente aos termos processuais recuperados do corpus.

5. Conclusões e trabalhos futuros

O presente artigo apresentou um estudo de caso no domínio da hemoterapia sobre a extração de termos médicos de um corpus, criado nesta área, que foram utilizados como classes de uma ontologia sobre os processos envolvidos na manipulação do sangue humano, em desenvolvimento no âmbito do *Blood Project*.

Embora se tenham atingido os propósitos aqui pretendidos, é importante ressaltar que, como pesquisa em andamento, ainda estão previstos passos como: (i) a validação das classes geradas para a ontologia, por parte de especialistas na área, com o propósito de assegurar maior representatividade dos termos; (ii) extração de termos compostos (bigramas, trigramas, etc.) para ontologia com base em técnicas estatísticas (exs: *índice de informação mútua, z-score)* e com uso de métodos de inferência; (iii) criação e processamento de um novo corpus na área de hemoterapia, que englobe as publicações científicas mais recentes na área. Tais passos são necessários para produzir resultados qualitativos mais consistentes e garantir maior qualidade à abordagem.

Como consideração final, acredita-se que as técnicas de PLN, de uma maneira geral, têm muito a contribuir com o processamento de grandes volumes de informações, tornando-o mais ágil e reduzindo drasticamente o tempo gasto por profissionais que

desempenham tarefas nesse contexto, tal como os curadores. No entanto, consideramos também que a intervenção humana é indispensável em algumas etapas da extração automática de termos para ontologias usando ferramentas de PLN, já que, atualmente, elas ainda não são capazes de tomar decisões próprias de especialistas humanos, baseando-se, exclusivamente, em informações linguísticas e estatísticas.

Referências

AABB – Advancing Transfusion and Cellular Therapies Worlwide [site] (2012). AABB ©. Disponível em: http://www.aabb.org/Pages/Homepage.aspx.

Almeida, M. B.; Teixeira, L. M. D.; Coelho, K. C.; Souza, R. R. (2010) "Relações semânticas em ontologias: estudo de caso do *Blood Project"*. *Liinc em Revista*, v.6, n.2, setembro, Rio de Janeiro, p. 384- 410.

Biber, D. Representativeness in Corpus Design. (1993) *Literary and Linguistic Computing*, vol. 8, n. 4.

Buitelaar, P.; Cimiano, P.; Magnini, B. (2005) "Ontology learning from text: An overview". In: Buitelaar, P.; Cimiano, P.; Magnini, B. (Ed.). *Ontology learning from text: Methods, evaluation and applications*, v.123 of Frontiers in Artificial Intelligence and Applications. IOS Press.

Ciminano, P.; Völker, J. (2005) *A framework for ontology learning and data-driven change discovery*. Institute AIFB, University of Karlsruhe, Germany.

Friedman, C.; Hripcsak G. (1999) "Natural language processing and its future in medicine". *Academic Medicine*, vol. 74, n. 8, Agosto.

Grenon, P.; Smith, B. (2004) "SNAP and SPAN: Towards Dynamic Spatial". *Spatial Cognition e Computation*, v.4, n.1, p. 69-104.

Leech, G. (1993) "Corpus annotation schemes". *Literary and Linguistic Computing* 8(4): 275-81.

McEnery, T.; Wilson, A. (2011) *Corpus Linguistics: an introduction*. Edinburgh: Edinburgh University Press, Second Edition.

Smith, B.; Welty, C (2001). "Ontology: Towards a new synthesis". In: Smith, B.; Welty, C. (Eds.). *Proceedings of the International Conference on Formal Ontology in Information Systems*. New York: ACM Press, p. 3–9.

Smith, B.; Ceusters, W.; Klagges, B.; Köhler, J.; Kumar, A.; Lomax, J.; Mungall, C.; Neuhaus, F.; Rector, A. L.; Rosse, C. (2005) "Relations in biomedical ontologies". *Genome Biology*, 6, R46, abr.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*. Philadelphia: John Benjamins BV. p. 47-64.

Wächter, T.; Schroeder, M. (2010) "Semi-automated ontology generation within OBO-Edit". *BioInformatics*, vol. 26, p. 88-96.

Winnenburg, R.; Wächter, T., Plake, C.; Doms, A.; Schroeder, M. (2008) "Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?" *Briefings in Bioinformatics*, vol. 8, n. 6, p. 466-478.