SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL
3ºONTOBRAS
30 e 31 de agosto de 2010 - Florianópolis/SC

Enriquecimento de Ontologias: uma Abordagem para Extração de Conhecimento do Campo Definição

Miguel G. P. Carvalho (UFRJ) miguelgabriel@ufrj.br

Vanessa Braganholo (UFRJ) braganholo@dcc.ufrj.br

Maria L. M. Campos (UFRJ) mluiza@nce.ufrj.br

Maria L. A. Campos (UFF) maria.almeida@cnpq.br

Resumo: Ontologias estão presentes em diversas áreas do conhecimento. Em especial, nesse trabalho estudamos as ontologias do domínio biomédico. Embora amplamente utilizadas pelas áreas relacionadas à bioinformática, essas ontologias possuem diversos problemas em sua construção e estruturação. Alguns desses problemas podem ser minimizados, com a possibilidade de extração de informações implícitas do campo definição. O foco principal desse trabalho é o enriquecimento das ontologias através de informações extraídas desse campo. Para alcançar esse objetivo é proposta uma abordagem de processamento textual baseada em quatros enfoques: enfoque baseado em aprendizado de máquina, utilizado para classificar frases e reconhecer padrões; enfoque baseado em dicionário, utilizado diretamente para encontrar termos e verbos (que podem indicar relações) no campo definição; enfoque baseado em regras, utilizado para identificar e extrair padrões; e, finalmente, o enfoque baseado em estatística para reconhecimento de termos (conceitos) relevantes.

Palavras-chave: Ontologias; Extração de informação; Informações implícitas

1. Introdução

Segundo Guarino (1998), "ontologia é um artefato de engenharia, constituído por um vocabulário específico usado para descrever uma certa realidade, mais um conjunto de pressupostos explícitos sobre o significado pretendido das palavras do vocabulário". Ontologias estão presentes em diversas áreas do conhecimento. Em especial, nesse trabalho, estudamos as ontologias biomédicas que são amplamente utilizadas pela área de bioinformática, onde técnicas e conhecimento de computação são empregados para o armazenamento, a análise e a geração de dados biológicos (Campos, 2009). Na área de bioinformática, as ontologias e outros esquemas terminológicos vêm sendo usados, por exemplo, como mecanismo para representação de conhecimento do domínio e como importante referência para anotação de dados genômicos (Belloze *et al.*, 2007).

Nesse contexto, o consórcio OBO (2010) é o mantenedor do mais importante repositório de terminologias. O objetivo desse consórcio é a criação de princípios para o

desenvolvimento de ontologias ortogonais e interoperáveis para uso compartilhado entre diversos domínios biológicos e médicos.

Um dos vocabulários mais utilizados hoje na OBO é a Gene Ontology (2010), a qual se divide em três ramos: componentes celulares, funções moleculares e processos biológicos (Lewis, 2004). Embora amplamente utilizada, essa ontologia e diversas outras do consórcio OBO ainda apresentam problemas na forma como são construídas e estruturadas. Algumas deficiências encontradas nas ontologias discutidas na literatura (Smith, Williams e Schulze-Kremer, 2003; Wroe *et al.*, 2003; Smith, Köhler, Kumar, 2004; Smith e Kumar, 2004; Ogren *et al.*, 2004; Ogren , Cohen e Hunter, 2005; Smith *et al.*, 2005; Schulz , Kumar e Bittner, 2006; Kohler *et al.*, 2006) e levantadas no decorrer do trabalho são: (i) **problemas na hierarquia:** deficiência na forma como os conceitos são estruturados e/ou omissão de termos importantes na hierarquia; (ii) **problemas nas relações:** poucas relações disponíveis para descrição dos relacionamentos entre os termos e para expressar o conhecimento do domínio das ontologias, relações empregadas de forma equivocada e/ou omissão de relações aparentemente óbvias; (iii) **problemas na descrição/definição dos termos:** informações implícitas inconsistentes com o conhecimento explicitado na ontologia e/ou falta de padronização para criação de descrições/definições; (iv) **problemas de contexto:** falta de regras para determinar se um dado conceito está ou não presente em uma determinada ontologia, ou seja, falta de descritores para definir que conceitos são expressos nas ontologias, falta de uma padronização mais formal para a construção das ontologias, problemas com nomenclatura dos termos, fazendo com que termos idênticos representem conceitos diferentes e termos diferentes representem o mesmo conceito, falta de documentação das ontologias e/ou falta de alinhamento entre as ontologias que algumas vezes possuem informações repetidas em relação a outras ontologias, não se relacionando explicitamente.

Grande parte dos problemas das ontologias interfere diretamente no processo de anotação e alinhamento, dificultando esse processo, levando em alguns casos a anotações e alinhamentos equivocados e incompletos (Smith *et al.*, 2005; Smith e Rosse, 2004).

A partir da lista de problemas apresentada, neste trabalho foi proposta uma abordagem para tirar vantagem do conhecimento contido no campo definição. A proposta desse

trabalho é utilizar técnicas de mineração de texto para estudar e extrair informações novas e úteis do campo definição das ontologias biomédicas, com o propósito de enriquecer essas ontologias. Para alcançar esse objetivo foi desenvolvida uma abordagem de processamento textual baseada em quatros enfoques: enfoque baseado em aprendizado de máquina, utilizado para classificar sentenças e reconhecer padrões; enfoque baseado em dicionário, utilizado diretamente para encontrar termos e verbos (que podem indicar relações) no campo definição; enfoque baseado em regras, utilizado para identificar e extrair padrões através de expressões regulares; e, finalmente, o enfoque baseado em estatística para reconhecimento de termos (conceitos) relevantes. Como exemplo, no campo definição associado ao termo **Mitochondria** (GO:0005739) tem-se: *"A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration"*[1]. A partir dessa definição, é possível extrair três relações: **is_a** *semiautonomous, self replicating organelle*; **occurs in** *varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells*; **the site of** *tissue respiration*, que eventualmente podem vir a complementar ou servir para tornar consistentes as relações já expressas explicitamente na estrutura da ontologia.

O restante deste artigo está organizado da seguinte maneira. Na seção 2 é feito um estudo do campo de definição. Na seção 3 são apresentados conceitos sobre extração de informação, importantes para o entendimento deste trabalho. Já na seção 4 é proposta a abordagem de processamento textual para tratamento do campo definição das ontologias. A seção 5 discute alguns resultados preliminares. Por fim, a seção 6 apresenta as considerações finais.

2. Ontologias e Definições

Segundo Corcho, Fernández-López e Gómez-Pérez (2003) as ontologias podem ser classificadas quanto ao seu grau de formalismo em ontologias pesadas (heavyweight) e ontologias leves (lightweight). Ambos os tipos de ontologias possuem conceitos, estrutura taxonômica dos conceitos, relações entre os conceitos e propriedades que descrevem os conceitos. Entretanto, nas ontologias pesadas adicionam-se axiomas e

[1] Neste artigo, optou-se por manter termos e definições segundo seu original em inglês, para que estes se mantivessem o mais possível fiéis à sua redação original.

restrições com a finalidade delimitar a semântica pretendida para o domínio. Embora as ontologias nem sempre apresentem os mesmos componentes, existem características e componentes básicos comuns presentes em grande parte delas. De forma geral, as ontologias são formadas por termos, relações e definições (Sales, Campos e Gomes, 2008). Os termos representam conceitos. Esses conceitos estão relacionados a outros através de relações e são descritos em maior detalhe através do campo definição.

O campo definição, estudado amplamente nesse trabalho, tem suma importância nas ontologias, pois é através desse campo que as pessoas compreendem os conceitos sobre os termos e conceitos da ontologia. Diversos outros trabalhos (Michel, Mejino e Rosse, 2001; Smith e Rosse, 2004; Köhler *et al.*, 2006) abordam a importância de uma definição de termos bem estruturada e formalizada, permitindo a compreensão dos conceitos por seres humanos e a inferência de conceitos e descoberta de conhecimento através de um raciocínio automático feito por máquinas.

Neste contexto, problemas na definição de conceitos têm sido objeto de diversos estudos ao longo dos anos. Em 1982, por exemplo, o *Groupe Interdisciplinaire de Recherche Scientifique et Appliquée en Terminologie* (GIRSTERM) já discutia as especificidades da definição em terminologia (por exemplo: O que definir? Como definir? Por que definir? A sinonímia dos termos se compara à sinonímia das palavras?) através de um Colóquio Internacional de Terminologia sob o tema "Problemas da definição e da sinonímia em terminologia" (Gomes e Campos, 1994).

Nessa linha de pensamento, Dahlberg (1981, p.17) apresenta o conceito de definição como: “a equivalência entre um *definiendum* (o que deve ser definido) e um *definiens* (como algo deve ser definido) com o propósito de delimitar o entendimento do *definiendum* em qualquer caso de comunicação. Em seu estudo, Dahlberg (1978) afirma que a estrutura de *definiens* mais conhecida e também a mais antiga é a *genus proximum et differentia specifica*. Essa estrutura de *definiens* relaciona um dado conceito a um conceito mais abrangente ou a um conceito similar (*genus proximum*), possibilitando com isso que o conceito seja relacionado com algo conhecido. Com certa freqüência, também se verifica nas ontologias o emprego da estrutura de *definiens*, *genus* supremum. Esses *definiens* geralmente denotam formas especificadoras, que designam “propriedade”, “categoria”, tendo o mesmo princípio de uma relação hierárquica (Dahlberg, 1978; Gomes e Campos, 1994). Nas ontologias biomédicas o *genus*

supremum é verificado com mais freqüência que o *genus proximum*, como por exemplo, em definições que começam por "uma reação química que...". Esse tipo de definição mostra uma relação hierárquica do termo descrito com a categoria de reação química.

Embora existam padrões para categorizar as definições e os conceitos, não existe uma fórmula única para trabalhá-las nos diferentes contextos. Em cada caso, deve-se tentar identificar os modelos mais apropriados para trabalhar/construir a definição, ou seja, identificar os padrões de enunciados definitórios em cada contexto. Através dessa análise, pode-se tirar vantagem da exploração deste campo com vistas à extração de informação e enriquecimento e busca de consistência nas ontologias.

3. Extração de Informação

Extração da informação é a aplicação de processamento de linguagem natural, com o objetivo de extrair informações do corpus. O processamento da linguagem natural é o conjunto de abordagens, mecanismos e técnicas para analisar o corpus escrito em linguagem natural (Ananiadou e McNaught, 2006). Com a finalidade de extrair informações desse corpus, são utilizados diversos enfoques para extração de informação. Vários destes enfoques fazem marcações no texto para identificar informações de interesse, como por exemplo: verbos, substantivos, etc. Os principais enfoques utilizados para extração de informações no domínio biomédico são (Ananiadou e McNaught, 2006):

Enfoque baseado em dicionário: utiliza um dicionário próprio para marcação de conceitos relevantes no corpus.

Enfoque baseado em regras: utiliza padrões de formação. Esse enfoque define padrões de formação para certo domínio e procura conceitos relevantes através desse padrão definido, utilizando-se para isso de análise ortográfica, léxica e morfossintática no corpus.

Enfoque baseado em aprendizagem de máquina: possui alto nível de prognosticar (prever) os termos e conceitos relevantes no corpus uma vez que infere conhecimento a partir do treinamento que foi recebido.

Enfoque baseado em estatística: baseia-se em diferentes distribuições estatísticas para inferência de conhecimento.

A Tabela 1 faz uma comparação entre esses enfoques mostrando suas vantagens e desvantagens. Nesse trabalho, para desfrutar das vantagens de cada um desses enfoques, será utilizada todos eles em diferentes etapas da abordagem proposta.

Tabela 1. Comparação entre os enfoques de extração da informação

Enfoque	Vantagem	Desvantagem
Dicionário	Fácil de ser implementado. Possui casamento (*matching*) perfeito.	Não identifica novos neologismos. Limitada as palavras do dicionário.
Regras	Melhor precisão.	Definição das regras (requer tempo).
Aprendizado de Máquina	Alto nível de prognosticar os termos.	Dependente do treinamento recebido. Deve ser treinado com um bom conjunto amostral.
Estatísticas	Alto nível de prognosticar os termos.	Definir medidas adequadas para aplicação no corpus.

Fonte: Ananiadou e McNaught, 2006

4. Abordagem de Processamento

A abordagem de processamento do campo definição das ontologias biomédicas proposta nesse trabalho é composta de 5 macro-etapas, representadas na Figura 1, que podem ser realizadas de forma semi-automática.

Figura 1: Abordagem de processamento.

A etapa de extração e transformação do corpus tem como função preparar o corpus que será utilizado para extração de informação. Nessa etapa as ontologias são transformadas em um arquivo XML que contém somente o nome dos termos e as definições.

A etapa de tratamento das definições aplica metodologias de tratamento de texto nas definições. Nos experimentos realizados, utilizou-se o GATE (*General Architecture for Text Engineering*) (Cunningham *et al*, 2002) e o NLTK (2010) como apoio computacional para tratar as definições (tokenização das sentenças; tokenização de palavras; radicalização (*stemming*); remoção de *stop words*; marcação do texto em categorias (*part-of-speech tagging*)) para serem usadas posteriormente. Os processos de radicalização e remoção de *stop words* são opcionais, uma vez que sua aplicação, em alguns casos, pode prejudicar a etapa de extração de informação.

A etapa de categorização das definições utiliza o **enfoque baseado em aprendizado de máquina** que classifica as definições em padronizadas (ex.: definições padronizadas da Gene Ontology) e não padronizadas e também separa definições com conceitos relevantes e não relevantes.

A etapa de extração de informação tem por objetivo a marcação e extração do corpus em palavras ou conjunto de sentenças relevantes que podem indicar: relacionamentos entre termos (conceitos). Essa marcação é utilizada para uma análise posterior das informações extraídas. Essa etapa utiliza-se de três enfoques: **enfoque baseado em dicionário**, utilizado diretamente para encontrar termos e verbos (que podem indicar relações) no campo definição, **enfoque baseado em regras** utilizado para extração de informações de padrões pré-definidos na ontologia (como por exemplo, os padrões da GO selecionados pelo enfoque de aprendizagem de máquina) através de expressões regulares que permitam que sejam inferidas relações entre termos e o **enfoque baseado em estatística**, utilizado para reconhecimento de termos (conceitos) e verbos relevantes através de cálculos de relevância no contexto em que são inseridos.

A etapa de análise das informações extraídas tem por objetivo fazer um estudo das informações extraídas, como por exemplo, validação das relações entre conceitos extraídos e análise da hierarquia dos termos com base nessas novas relações (verificando a compatibilidade com a hierarquia original da ontologia). Essa etapa precisa de um especialista da área para validação das informações extraídas.

A abordagem proposta nesse trabalho está em fase de implementação no mecanismo batizado de EI-Onto. O EI-Onto trabalha por sobre o framework GATE e sobre o mecanismo NLTK para o tratamento e extração de conhecimento do campo definição das ontologias biomédicas.

5. Resultados Preliminares

Embora esta abordagem ainda esteja em fase de implementação no mecanismo Ei-Onto, já foi possível, com sua aplicação na ontologia Gene Ontology e com um estudo do campo definição, a constatação de alguns fatos que merecem ser estudados mais a fundo.

1) O campo definição contém conhecimento que pode ser usado para enriquecer a ontologia. Ex.: No termo **Microbody Part** (GO:0044438) cuja definição é *"Any constituent part of a microbody, a cytoplasmic organelle, spherical or oval in shape, that is bounded by a single membrane and contains oxidative enzymes, especially those utilizing hydrogen peroxide (H2O2)"*, é possível extrair três relações: **part of** *a microbody, a cytoplasmic organelle, spherical or oval in shape*; ***is bounded by*** *a single membrane*; ***contains*** *oxidative enzymes.* O grande desafio nesse caso é avaliar quais relações precisam ser incorporadas explicitamente na ontologia.

2) Muitos nomes de termos aparentemente exprimem uma relação com outros termos com os quais não possuem relação explícita na ontologia. Ex.: *host cell cytoplasm* (GO:0030430) e *cytoplasm* (GO:0005737).

3) Muitos termos citados na definição não possuem relação com o termo que está sendo descrito/definido. Ex.: O termo **Transcription factor TFIIIB-alpha complex** (GO:0034732) cita o termo **Transcription factor TFIIIB-beta complex** (GO:34733) em sua definição: "*A transcription factor TFIIIB-beta complex that contains the TATA-binding protein (TBP), B"[...]"*, porém não possui qualquer relação explícita com o primeiro.

4) A Gene Ontology possui alguns padrões definitórios que possibilitam a extração de informação relevante como novos termos e relacionamentos. Ex.: [enzima][2] *activity* que possui padrão de definição: *Catalysis of reaction*: [reação catalisada pela enzima][2]. Através dessa definição é possível verificar que todos os termos desse grupo devem possuir como um ancestral o termo "*Catalysis*".

Os resultados preliminares dessa seção foram trabalhados usando a API do Java Jena (2010) para navegação em ontologias e a interface da GO AmiGO (Gene Ontology, 2010).

6. Considerações Finais

Este artigo apresenta algumas considerações importantes a respeito do campo definição das ontologias biomédicas que merecem ser estudadas com mais detalhes. Uma questão

[2] O texto entre colchetes deve ser substituído

abordada nesse trabalho, que merece um aprofundamento maior, é a verificação da consonância do campo definição com a estrutura sistemática classificatória da ontologia. Na aplicação dessa abordagem pode ser verificada a omissão de termos que deveriam pertencer à estrutura hierárquica do termo, pois aparecem em sua definição, mas que não são encontrados nesta hierarquia. Essa inconsistência pode representar problemas na interpretação de um termo ou eventuais falhas na estrutura classificatória originalmente criada.

A segunda questão é a de que é possível o enriquecimento do conhecimento das ontologias através da incorporação de novas relações implícitas no campo definição. Porém, não podemos afirmar que isto deva sempre resultar em complementação das relações na ontologia, pois sua relevância deve ser avaliada para evitar um aumento excessivo de complexidade na representação.

Além da implementação da abordagem proposta e aprofundamento dos estudos nas questões levantadas, esse trabalho, no decorrer de seu desenvolvimento, ainda tem por objetivo criar uma abordagem de extração de informações através da combinação dos quatro enfoques citados (aprendizagem de máquina, regras, dicionário, estatísticas) que embora inicialmente seja desenvolvida para aplicação em ontologias do domínio biomédico, possa futuramente, com algumas modificações, ser aplicada a outros domínios.

Referências

ANANIADOU, S.; McNAUGHT, J. Text Mining for Biology and Biomedicine. Norwood: Artech House, 286p, 2006.

BELLOZE, K. T.; DÁVILA, A.; ARAUJO, R. M.; CAVALCANTI, M. C. Ampliando o Uso Colaborativo de Ontologias em Processos de Anotação Genômica. In: E-Science Workshop-SBBD/SBES, p. 71-80. João Pessoa, Paraíba, Brazil, 2007.

CAMPOS, M.L.A. Aspectos semânticos da compatibilização terminológica entre ontologias no campo da Bioinformática In: X Encontro Nacional de Pesquisa da ANCIB - ENANCIB, João Pessoa, 2009.

CORCHO, O.; FERNÁNDEZ-LÓPEZ, M; GÓMEZ-PÉREZ, A. Methodologies, Tools and Languages for Building Ontologies. Where is Their Meeting Point? Data & Knowledge Engineering. p.41–64, 2003.

CUNNINGHAM, H.; MAYNARD, D.; BONTCHEVA, K.; TABLAN, V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002

DAHLBERG, I. A Referent-Oriented, Analytical Concept Theory for INTERCONCEPT. In: International Classification. v.5, n.3, p.142-150, 1978.

DAHLBERG, I. Conceptual Definitions for INTERCONCEPT. In: International Classification. v.8, n.1, p.16-22, 1981.

GENE ONTOLOGY. AmiGO. Disponível em:< http://amigo.geneontology.org>. Acessado em: 1-jun-2010.

GENE ONTOLOGY. Disponível em:< http://www.geneontology.org/>. Acessado em: 1-mai-2010.

GOMES, H.; CAMPOS, M.L.A. Contribuição da Teoria da Terminologia na Elaboração de Hiperdocumentos com Função de Tesauro. In: IV Simpósio Iberoamericano de Terminología RITerm, Buenos Aires, Argentina, 2004

GUARINO,N. Formal Ontology and Information Systems. In: Proceedings of Formal Ontology in Information Systems (FOIS'98),Trento, Italy, p. 3-15, 1998.

JENA. Disponível em: < http://jena.sourceforge.net/ >. Acessado em: 2-jun-2010.

KÖHLER, J.; MUNN, K.; RUEGG, A.; SKUSA, A.; SMITH, B. Quality Control for Terms and Definitions in Ontologies and Taxonomies. BMC Bioinformatics, 2006.

LEWIS, S.E. Gene Ontology: Looking Backwards and Forwards. Genome Biology. v.6, n.1, p.1-4, 2004.

MICHAEL, J.; MEJINO-JR., J.L.V; ROSSE, C. The Role of Definitions in Biomedical Concept Representation. In: Proceedings of AMIA Symposium. p.463-467, 2001.

NLTK. Disponível em:< http://www.nltk.org/>. Acessado em: 1-mai-2010.

OBO: Open Biomedical Ontologies. Disponível:< http://www.obofoundry.org/> Acessado em: 20-mai-2010.

OGREN, P.V.; COHEN, K.B.; ACQUAAH-MENSAH, G.K.; EBERLEIN, J.; HUNTER, L. The Compositional Structure of Gene Ontology Terms. In: Proceedings of the Pacific Symposium on Biocomputing. v.9, p.214-225, 2004.

OGREN, P.V.; COHEN, K.B.; HUNTER, L. Implications of Compositionality in the Gene Ontology for Its Curation and Usage. In: Proceedings of the Pacific Symposium on Biocomputing. v.10, p.174-185, 2005.

SALES, L.F.; CAMPOS, M.L.A; GOMES, H.E. Ontologias de Domínio: Um Estudo das Relações Conceituais. Perspectivas em Ciência da Informação, Belo Horizonte. v.13, n.2, 2008.

SCHULZ, S.; KUMAR, A.; BITTNER, T. Biomedical Ontologies: What part-of is and isn't. Journal of Biomedical Informatics. v.39, n. 3, p. 350-361, 2006.

SMITH, B.; CEUSTERS, W.; KLAGGES, B.; KÖHLER, J.; KUMAR, A.; LOMAX, J.; MUNGALL, C.; NEUHAUS, F.; RECTOR, A.L.; ROSSE, C. Relations in Biomedical Ontologies. Genome Biology. v.6, n.5, p.1-15, 2005

SMITH, B.; KÖHLER, J.; KUMAR, A. On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. In: Database Integration in the Life Sciences, p.1-17, 2004.

SMITH, B.; KUMAR, A. On Controlled Vocabularies in Bioinformatics: A Case Study in the Gene Ontology. In: Drug Discovery Today: BIOSILICO. v.2, n.6, p.246-252, 2004.

SMITH, B.; ROSSE, C. The Role of Foundational Relations in the Alignment of Biomedical Ontologies. In: World Congress on Medical Informatics (MEDINFO), Amsterdam: IOS Press, p.444-448, 2004.

SMITH, B.; WILLIAMS, J.; SCHULZE-KREMER, S. The Ontology of the Gene Ontology. In: Proceedings of AMIA Symposium. p.609-613, 2003.

WROE, C. J.; STEVENS, R.; GOBLE, C.A.; ASHBURNER, M. Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL. In: Pacific Symposium on Biocomputing. v.8, p.624-635, 2003.