Construção de Modelos Conceituais a Partir de Textos com Apoio de Tipos Semânticos

Felipe Leão, Thaíssa Diirr, Fernanda Baião, Kate Revoredo

NP2Tec – Núcleo de Pesquisa e Prática em Tecnologia
Departamento de Informática Aplicada
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Av. Pasteur, 296 Urca – Rio de Janeiro – Brasil

{felipe.leao, thaissa.medeiros, fernanda.baiao, katerevoredo}@uniriotec.br

Abstract. *The conceptual modeling process involves the understanding of domain concepts and its representation in diagrams. The knowledge about the domain can be obtained from various sources, most of them being based on natural language. Therefore, the automatically creation of conceptual models becomes interesting. This paper presents an analysis of the feasibility of automating a linguistic approach with semantic focus to allow the construction of ontologies from natural language texts.*

Resumo. *O processo de modelagem conceitual envolve a compreensão de conceitos do domínio em questão e sua posterior representação em diagramas. O conhecimento sobre o domínio pode ser obtido a partir de diversas fontes, sendo a maior parte delas baseadas em linguagem natural. Portanto, a criação de modelos conceituais de forma automática se mostra interessante. Este trabalho apresenta uma análise da viabilidade de se automatizar uma abordagem linguística com foco semântico capaz permitir a construção de ontologias a partir de textos em linguagem natural.*

1. Introdução

A modelagem conceitual constitui uma das principais fases na construção de um sistema de informação. Em essência, tal fase busca definir os conceitos envolvidos na descrição do problema, sendo o seu resultado um substrato de grande valor para a boa compreensão geral dos requisitos e auxílio para as etapas seguintes na modelagem de qualquer solução computacional. O processo envolve a compreensão de conceitos relacionados ao domínio em questão e a posterior representação desses conceitos em modelos, como por exemplo ontologias.

Por outro lado, a transmissão de conhecimento do especialista no domínio para o modelador durante a construção de uma ontologia é geralmente feita através da troca de informações em textos (transcrições de entrevistas, documentos regulatórios, entre outros). Para obter uma representação fiel do domínio modelado com qualidade semântica, é necessário, além de um metamodelo representativo, que o modelador interprete corretamente os conceitos do domínio e não influencie sua compreensão com experiências e conhecimentos anteriores [Castro *et al*., 2011].

Diversos trabalhos propuseram abordagens para modelagem conceitual de dados através de ontologias com base em conceitos linguísticos, entre eles Castro *et al*. (2011) propôs um método baseado em linguística e foco semântico. O processo é naturalmente

composto por duas atividades principais, (i) a aquisição de conceitos sendo utilizados no domínio a ser modelado e (ii) a representação de tais conceitos através da linguagem de modelagem OntoUML [Guizzardi, 2011]. Apesar de esse método auxiliar a produção de ontologias como modelos conceituais de dados com qualidade semântica, a falta de suporte automatizado é um ponto negativo, uma vez que as definições formais de uma língua podem ser de grande complexidade. É interessante então buscar meios que viabilizem a automatização da abordagem.

Este trabalho se propõe a analisar a aplicação de técnicas automatizadas para reconhecimento sintático e semântico à abordagem de Castro et al. (2011), auxiliando e agilizando a realização dos passos que a compõem. Técnicas utilizadas atualmente para realização de processamento de linguagem natural como *POS Tagging* e *Semantic Tagging* foram pesquisadas.

O restante do trabalho está organizado da seguinte forma: A seção 2 apresenta a abordagem linguística para modelagem com foco semântico, incluindo a linguagem OntoUML. A seção 3 descreve as técnicas analisadas e as duas abordagens consideradas. A seção 4 expõe considerações finais sobre o trabalho e trabalhos futuros.

2. Abordagem Linguística para Modelagem Conceitual com Foco Semântico

Castro et al. (2011) propôs uma abordagem linguística para modelagem conceitual que parte da linguagem natural para construção de ontologias em OntoUML. Esta seção explicita os passos propostos para obtenção dos modelos, além da própria linguagem OntoUML.

2.1. OntoUML

A OntoUML é uma linguagem de modelagem conceitual ontologicamente bem fundamentada, que pode apoiar a fase de análise de domínio na engenharia de sistemas de informação. Ela foi proposta como um perfil da UML 2.0 e incorpora axiomas e distinções ontológicas presentes na ontologia de fundamentação UFO (*Unified Foundation Ontology*) [Guizzardi, 2011].

Os elementos da UFO podem ser *Universal* (tipos que determinam coleções de conceitos que se referem a coisas ou seres) ou *Individual* (instâncias de tais coleções). Essas classes se subdividem em outras que, por sua vez, também possuem sub-hierarquias [Guizzardi, 2005 apud Castro et al., 2011]. Na hierarquia de *Universal*, por exemplo, poderíamos citar a classe *Substantial Universal* que corresponde a propriedades intrínsecas (não relacionais) que determinam classes ou coleções de seres ou coisas materiais e mantêm sua identidade, mesmo passando por mudanças.

As classes da linguagem de modelagem conceitual OntoUML são especializações das classes abstratas da UFO, herdando meta-propriedades e/ou restrições. Alguns desses construtos são [Guizzardi, 2005; Guizzard, 2011; e Castro et al., 2010]:

- *Kind* (tipo): seres complexos ou coisas relacionalmente independentes claramente identificados. (e.g. animais, plantas, time de futebol).
- *Quantity* (quantidade): classes de substâncias de massa (e.g. água).

- *Collective* (coletivo): coleções ou coisas vistas e percebidas como uma estrutura uniforme (e.g. floresta, baralho de cartas).
- *Phase* (fase): estágios ou fases na existência de um ser (e.g. lagarta e borboleta são partições - *phase* - de um lepidópteros - *kind*).

2.2. Modelagem Conceitual com Foco Semântico

O processo de modelagem de dados conceitual é similar a uma atividade de tradução, pois consiste em entender os conceitos representados em uma linguagem natural e, então, representar esses mesmos conceitos em uma linguagem de modelagem [Castro *et al*., 2011]. Para realizar essa tradução é preciso compreender e comparar as linguagens, a fim de elaborar modelos conceituais (ontologias) semanticamente precisos que apresentem o mesmo significado do texto em linguagem natural [Castro *et al*., 2010 e 2011].

O vocabulário de qualquer linguagem natural é dividido em classes de palavras que podem ser fechadas ou abertas. Classes abertas são as que carregam carga semântica (substantivos, adjetivos, verbos e advérbios) e, por isso, são o foco da modelagem conceitual. As classes de palavras podem ainda ser especializadas em tipos semânticos propostos por Dixon (2005). Alguns destes tipos especializam os substantivos de referência concreta, que têm maior importância para modelagem conceitual por nomearem seres e coisas. São eles:

- *Animate* (animados): abrangem os animais não humanos;
- *Human* (humanos): referentes a seres humanos. São subdivididos em *Kin* (e.g. filho, pai), *Rank* (e.g. goleiro, professor) e *Social group* (e.g. companhia).
- *Parts* (partes): são partes de outras coisas ou seres, incluindo as partes corpóreas.
- *Inanimate* (inanimados): subdivididos em *Artifacts* (e.g. livro), *Flora* (e.g. árvore), *Celestial and Weather* (e.g. lua, vento) e *Environment* (e.g. água).

A partir do entendimento das propriedades das linguagens naturais e da OntoUML, Castro *et al*. (2010) definiram mapeamentos de tipos semânticos relacionados a substantivos, verbos e adjetivos para construtos da OntoUML. A Tabela 1 apresenta os construtos da OntoUML relacionados aos tipos semânticos de substantivos concretos.

Tabela 1 - Mapeamento de Tipos Semânticos de Substantivos Concretos [Dixon, 2005] para OntoUML. Fonte: Castro et al (2010)

Tipos Semânticos	Construtos OntoUML
Animate, Human, parts, Inanimate, Inanimate/Artifacts, Inanimate/Celestial and Weather, Inanimate/ Flora	Kind
Human/Social Group	Collective ou Kind
Inanimate/Environment	Quantity
*Vários	Phase

Esses mapeamentos são utilizados na abordagem de modelagem conceitual proposta em Castro *et al*. [2010, 2011]. A abordagem parte de textos descritivos sobre o universo a ser modelado, produzidos pelos especialistas do domínio. Os passos que

compõem a abordagem estão listados a seguir [Castro *et al*., 2011], onde o mapeamento da Tabela 1 é aplicado no passo 5.

1. *Decomposição do texto em sentenças simples* (sentenças não interrogativas, na voz ativa e contendo somente uma oração);
2. *Levantamento e esclarecimento de dúvidas junto ao especialista do domínio;*
3. *Identificação dos signos do universo modelado* (sujeitos, verbos e objetos);
4. *Associação entre os signos identificados e os tipos semânticos correspondentes;*
5. *Mapeamento entre os tipos semânticos e os construtos da OntoUML;*
6. *Criação do modelo*.

3. Técnicas para Reconhecimento Sintático e Semântico

É possível observar que os passos 1, 3 e 5 da abordagem de Castro et al (2010) são diretamente automatizáveis, através de técnicas como *Tokenizing* e *Parsing* e os mapeamentos de tipos semânticos para OntoUML. O passo 2 consiste em uma tarefa não automatizável. Porém, para os passos 4 e 6, uma análise mais profunda se faz necessária. Optamos por focar no passo 4, por acreditarmos que esta tarefa é a que pode apresentar o maior desafio para a automatização do método como um todo.

A fim de auxiliar e agilizar a realização dos passos que compõe a abordagem de Castro et al. (2011), seria interessante o uso de algum procedimento automatizado. Em seu trabalho, Castro chegou a analisar uma ferramenta para anotação semântica, o Palavras (http://visl.sdu.dk/visl/pt/). Apesar de identificar classes gramaticais e funções sintáticas corretamente, ocorreram falhas quando se atingiu o nível semântico. Neste trabalho é feita uma análise mais abrangente sobre esta classe de ferramentas.

Inicialmente, analisamos ferramentas de processamento de linguagem natural que realizam anotação sintática de partes do discurso (*Part-of-Speech Tagging*), ou seja, o processo de classificar palavras morfologicamente de acordo com as classes gramaticais [Bird *et. al*., 2009]. Em seguida, buscamos abordagens que realizam algum tipo de anotação semântica (*semantic tagging*) para classificação das palavras, considerando informações de contexto.

3.1. Abordagens de *POS Tagging*

O estudo realizado consistiu em aplicar algoritmos de *POS tagging* a uma lista de sentenças simples tratadas. Foram testados diferentes *frameworks* como o NLTk (http://nltk.org), o Apache OpenNLP (http://opennlp.apache.org/) e o Stanford Parser/Tagger (http://nlp.stanford.edu/). O primeiro possibilitou a aplicação dos algoritmos de *POS-tagging* N-Gram (considerando suas variações Unigran, Bigram e Trigram) e Brill Tagger [Brill, 1992]. O segundo e o terceiro implementam o Modelo de Entropia Máxima (*Maximum Entropy Model* ou MaxEnt Model) [Ratnaparkhi, 1996].

3.1.1. Análise Prática dos Algoritmos de POS Tagging

Os três algoritmos (N-Gram, Brill e MaxEnt) foram aplicados a um mesmo conjunto de 56 sentenças já processadas e resultantes do estudo de caso de Castro et al. (2011) após decomposição em sentenças simples, divisão em orações, depassivização, separação em sentenças nucleares, e criação e esclarecimento de lista de dúvidas. O objetivo era analisar as anotações geradas pelos algoritmos para os substantivos identificados por Castro como os signos importantes do universo a ser modelado.

Os três algoritmos foram capazes de anotar corretamente a maior parte dos termos das sentenças, identificando suas classes gramaticais com precisão similar à indicada pelos seus criadores. Entretanto, o resultado dos *taggers* se mostrou tão limitado quanto o da ferramenta Palavras, uma vez que os tipos semânticos não puderam ser especificados. Os corpora não contêm, em seu conjunto de *tags* possíveis, indicações de tipos semânticos. A Figura 1 exemplifica a aplicação do Stanford Tagger, onde observa-se que, que mesmo capaz de identificar o substantivo "professor", o *tagger* que não faz qualquer indicação sobre este ser um *Human* e mais especificamente um *Rank*. Sem um corpus que indique tais classificações, a análise sobre a capacidade dos *taggers* em reconhecer o contexto semântico não se fez possível.

Sentença Original	***"O Professor responsável pela disciplina defere ou indefere os requerimentos de inscrição em disciplina isolada."***
Stanford Tagger	`[('O', [art]) ('professor', [n]) ('responsável', [adj]) ('pela', [v-fin]) ('disciplina', [n]) ('defere', [v-fin]) ('ou', [conj-c]) ('indefere', [v-fin]) ('os', [art]) ('requerimentos', [n]) ('de', [prp]) ('inscrição', [n]) ('em', [prp]) ('disciplina', [n]) ('isolada', [v-pcp]) ('.', [punc])]`

Figura 1 - Exemplo de *taggeamento* realizado pelo Stanford Tagger

3.2. Abordagens de *Semantic Tagging*

Uma vez observado que os *taggers* atualmente utilizados pela comunidade de NLP não eram capazes de identificar os tipos semânticos dos signos no texto, optou-se por estender a pesquisa para abordagens que considerassem a anotação semântica (*semantic tagging*). O que foi percebido com a pesquisa é que o termo "*semantic tagging"* tem sido utilizado para denotar diferentes técnicas em áreas de atuação distintas, desde as que de fato abordam o processamento de linguagem natural até aquelas que tangem outras tarefas como a de criação automática de taxonomias ou apoio a criação de *folksonomias*. Ainda que não tenham sido encontrados trabalhos que permitissem a automatização da identificação dos tipos semânticos propostos por Dixon (2005) a partir dos signos de um texto, alguns trabalhos podem indicar caminhos interessantes a serem seguidos na busca pelo desenvolvimento (ou adaptação) de um *tagger* semântico.

Alguns trabalhos como os de Mahar e Memon (2010) e Miller *et. al*. (1993), propõe a utilização de bases de dados léxicas como o WordNet para restringir a semântica de termos de acordo com os outros termos presentes nas sentenças, o que poderia auxiliar no desenvolvimento de um *tagger* semântico. Gelbukh e Kolesnikova (2012) também fazem uso do WordNet ao introduzirem o problema de se reconhecer, automaticamente, o sentido das palavras participantes de "*collocations"* (colocações), situação aonde ao se combinar duas palavras uma delas tem seu valor semântico alterado. Buitelaar (1997) propõe o léxico CoreLex para *tagging* semântico, capaz de classificar informações em um nível semântico utilizando conceitos compatíveis com os construtos da OntoUML utilizados na abordagem de Castro et al. (2011).

4. Considerações e Trabalhos Futuros

Este trabalho analisou técnicas utilizadas atualmente pela comunidade de processamento de linguagem natural observando o quanto elas podem apoiar um método de modelagem conceitual com foco semântico, especificamente o proposto por Castro et al. (2011). O objetivo deste método é elaborar uma ontologia como modelo

conceitual de domínio semanticamente expressivo. O método é baseado em linguística e gera um modelo conceitual bem fundamentado em OntoUML a partir de textos em linguagem natural. A aplicação de tipos semânticos possibilita o aumento no poder de expressão e diminui as chances de erros decorrentes do processo de modelagem.

Foram aplicadas técnicas de anotação gramatical através dos *frameworks* NLTk, OpenNLP e Stanford Tagger. Foi observado resultado similar ao descrito por Castro et al. (2011) quando a ferramenta Palavras foi analisada. Optou-se então por buscar soluções alternativas com foco em *tagging* semântico e alguns trabalhos relacionados ao tema foram revistos, com suas possíveis colaborações explicitadas.

Não foi encontrada uma solução para *tagging* semântico que possa ser diretamente aplicada ao problema aqui abordado, entretanto foi possível detectar tópicos que combinados e estendidos poderiam auxiliar na automatização do método de Castro et al. (2011). Outras abordagens como Desambiguação de Sentidos de Palavras (*Word Sense Desabiguition*) podem também ser de grande ajuda e deverão ser objetos de estudo em trabalhos futuros. É importante ressaltar que apesar das técnicas aqui relatadas não serem suficientes de maneira isolada para a automatização da abordagem, elas podem diminuir o número de falhas no processo de modelagem conceitual, uma vez que a análise de textos em linguagem natural pode ser uma tarefa complexa até mesmo quando os indivíduos trabalham com suas línguas maternas.

Agradecimentos

O primeiro autor agradece ao Programa CAPES/REUNI pela bolsa de estudos recebida.

Referências

Bird, S., Klein, E., Loper, E. (2009) “Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit”. O’Reilly Media.

Brill, E. (1992) “A simple rule-based part of speech tagger”. In Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155.

Buitelaar, P. (1997) “A Lexicon for Underspecified Semantic Tagging”. Proceedings of ANLP 97, SIGLEX Workshop.

Castro, L., Baião, F. A., Guizzardi, G. (2010) “A Linguistic Approach to Conceptual Modeling with Semantic Types and OntoUML”. EDOCW 2010: 215-224.

Castro, L., Baião, F. A., Guizzardi, G. (2011) “A Semantic Oriented Method for Conceptual Data Modeling in OntoUML Based on Linguistic Concepts”. ER 2011: 486-494.

Dixon, R. (2005) “A Semantic Approach to English Grammar”. Oxford University Press.

Gelbukh, A., e Kolesnikova, O. (2012) “Supervised Learning Algorithms Evaluation on Recognizing Semantic Types of Spanish verb-noun Collocations”. Computación y Sistemas.

Guizzardi, G. (2005) “Ontological Foundations for Structural Conceptual Models”. Ph.D. dissertation, University of Twente, Enschede, The Netherlands.

Guizzardi, G., das Graças, A. P., Guizzardi, R. S. S. (2011) “Design Patterns and Inductive Modeling Rules to Support the Construction of Ontologically Well-Founded Conceptual Models in OntoUML”. CAiSE Workshops 2011: 402-413.

Mahar, J. A., e Memon, G. Q. (2010) “Sindhi Part of Speech Tagging System using Wordnet”. International Journal of Computer Theory and Engineering, Vol. 2, No. 4, August, 2010

Miller, G. A., Leacock, C., Tengi, R., Bunker, R. T. (1993) “A Semantic concordance”. Proceedings of the workshop on Human Language Technology.

Ratnaparkhi, A. (1996) “A Maximum Entropy Model for Part-Of-Speech Tagging”. Department of Computer and Information Science – University of Pennsylvania.