Descoberta Automática de Relações Não-Taxonômicas a partir de Corpus em Língua Portuguesa

Vinicius H. Ferreira, Lucelene Lopes, Renata Vieira

PPGCC – FACIN – Porto Alegre – Brasil

vinihf@gmail.com, {lucelene.lopes,renata.vieira}@pucrs.br

Resumo. *A construção de ontologias é um processo complexo que compreende etapas como a extração de conceitos de domínio e a extração de relações taxonômicas e não-taxonômicas entre esses conceitos. A etapa de extração de relações não-taxonômicas é a mais negligenciada, especialmente para textos em língua portuguesa. Com isto, este trabalho apresenta uma proposta de extração de relações não-taxonômicas a partir de textos em português representados por uma lista de conceitos e informações contextuais automaticamente extraídos pela ferramenta ExATOlp.*

Abstract. *Ontology construction is a complex process composed by extraction tasks for domain concepts, as well as taxonomic and non-taxonomic relations among concepts. The extraction of non-taxonomic relations is the most neglected task, specially for Portuguese texts. Therefore, this paper presents a proposal for extracting non-taxonomic relations from Portuguese texts represented by a list of concepts and their contextual information extracted automatically by ExATOlp software tool.*

1. Introdução

A construção de ontologias de domínio [Gruber, 1993] é um ramo relevante da área de Processamento de Linguagem Natural (PLN). Tradicionalmente todo o processo de construção de ontologia é dependente de especialistas do domínio estudado. No entanto, esses especialistas são na maioria das vezes sobrecarregados pelo tamanho e complexidade dos dados e informações contidos no processo de construção de ontologias [Cimiano et al. 2006]. Com isto, busca-se métodos automáticos que reduzam a demanda de intervenção de especialistas [Maedche e Staab, 2000]. Dentre esses métodos, o presente trabalho interessa-se por aqueles que constroem ontologias a partir de *corpus*, que por sua vez é um conjunto de textos sobre um domínio específico [Biemann, 2005]. Dentre vários trabalhos [Maedche e Staab, 2000; Sanchéz e Moreno, 2008; Serra e Girardi, 2011; Villaverde et al. 2009; Schutz e Buitelaar, 2005], o processo proposto por Lopes (2012) gera, como uma das saídas possíveis, uma lista de conceitos e as informações contextuais da utilização de cada conceito no domínio. Dentre estas informações são identificadas a função sintática e os verbos aos quais o conceito se relaciona.

Para Maedche e Staab (2000), o processo de construção de ontologias contempla três etapas básicas: (i) extração de conceitos de domínio; (ii) extração de taxonomia; e (iii) extração de relações não-taxonômicas. A maior parte dos trabalhos

semelhantes coleta conceitos relevantes de um domínio e pode agrupá-los em uma hierarquia (taxonomia) utilizando métodos linguísticos e estatísticos [Lopes, 2012; Pantel e Lin, 2001; Chung, 2003; Brewster et al. 2003]. De acordo com Sanchéz e Moreno (2008), no processo de Aprendizagem de Ontologias, a fase de extração de relações não-taxonômicas tem sido reconhecida como a mais complexa e negligenciada [Maedche e Staab, 2000; Sanchéz e Moreno, 2008; Villaverde et al. 2009].

Diferente das relações taxonômicas, que contribuem na estruturação de um domínio e classificação de conceitos, as relações não-taxonômicas não estão relacionadas a hierarquia. Este tipo de relação acrescenta informações aos conceitos já encontrados, identificando relacionamentos entre eles [Guarino e Welty, 2002].

Identificar as relações não-taxonômicas é essencial para expressar as propriedades de classes e entidades de um domínio específico [Cimiano et al. 2006], representando as ações ou eventos que ocorrem entre os conceitos. Por exemplo, uma relação não-taxonômicas no campo do Direito, é a relação "representa" entre os conceitos "Advogado" e "Cliente" [Serra e Girardi, 2011], e no campo do Esporte, a relação "chuta" entre os conceitos "Jogador" e "Bola" [Schutz e Buitelaar, 2005].

De acordo com Serra e Girardi (2011), relações não-taxonômicas podem ser classificadas como independentes e dependentes de domínio. As relações independentes de domínio podem ser divididas em: (i) agregação, identificadas por relações "todo-parte"; e (ii) propriedade, identificadas por relações de posse ou composição. Relações dependentes de domínio são identificadas por termos específicos de um domínio.

O papel dos verbos como elemento de conexão central entre conceitos é inegável. Eles são responsáveis por especificar qual é a interação entre os participantes de uma ação ou evento, expressando a relação entre eles. Devido a isto os verbos tem sido muito utilizados para definir relações não-taxonômicas [Kavalec et al. 2004; Maedche e Staab, 2000; Sanchéz e Moreno, 2008; Schutz e Buitelaar, 2005].

Partindo disso, esse trabalho apresenta uma proposta de processo para extração de relações não-taxonômicas em textos na língua portuguesa. Diferente dos trabalhos similares de Sanchéz e Moreno (2008) e Brewster et al. (2003), aqui utiliza-se como fonte a lista de conceitos e informações contextuais geradas pelo ExATOlp que é uma ferramenta que implementa todas etapas do processo de extração de conceitos e contextos proposto por Lopes (2012).

2. Trabalhos Similares

Na literatura encontram-se trabalhos que propõe processos de extração de relações não-taxonômicas a partir de textos. A Tabela 1 apresenta uma síntese a respeito dos trabalhos similares ao apresentado neste artigo.

Pode-se observar através da Tabela 1, com exceção do trabalho de Sanchéz e Moreno (2008), todos os demais utilizam um conjunto de textos não-estruturados de um determinado domínio (*corpus* de domínio) como fonte para o processo de extração de relações não-taxonômicas. Dentre esses trabalhos, com exceção do trabalho de Maedche e Staab (2000), todos utilizam o verbo como elemento principal na identificação de relações não-taxonômicas. Dentro deste contexto, observou-se a constante ocorrência da manipulação da tripla *<conceito 1, conceito 2, verbo>* nos trabalhos apresentados.

Embora nem todos os trabalhos tenham seu foco em encontrar relações não-taxonômicas em corpus da língua inglesa, nenhum deles apresenta uma proposta para a língua portuguesa. Além disso, foi possível observar que na maioria dos trabalhos analisados as relações não-taxonômicas extraídas são dependentes do domínio, ou seja, a relação entre os conceitos é feita por termos específicos do domínio. Uma possível justificativa para isso é que nestes trabalhos os verbos que relacionam os conceitos são também usados como identificadores da relação.

Tabela 1. Comparação de trabalhos similares

Trabalhos Similares	Fonte de dados	Relação identificada por verbo?	Idioma	Tipos de relações não-taxonômicas extraídas
Maedech e Staab (2000)	*Corpus* de domínio	Não	Alemão	Dependente de domínio
Schutz e Buitelaar (2005)	*Corpus* de domínio	Sim	Inglês e Alemão	Dependente de domínio
Sanchéz e Moreno (2008)	Web	Sim	Inglês	Dependente de domínio
Villaverde et al. (2009)	*Corpus* de domínio, lista de candidatos a conceitos ou hierarquia de conceitos	Sim	Inglês	Dependente de domínio
Weichselbraun et al. (2009)	Ontologia e *corpus* de domínio	Sim	Inglês	Independente de domínio
Serra e Girardi (2011)	*Corpus* de domínio	Sim	Inglês	Dependente e independente de domínio

Na análise dos trabalhos similares foi possível observar que a maioria das propostas não se propõe a identificar relações não-taxonômicas de forma automática, mas sim sugerir relações para especialistas de domínio. Com isso, pode-se verificar que o papel dos engenheiros de ontologias e especialistas de domínio é importante na decisão final a respeito de relações não-taxonômicas [Serra e Girardi, 2011].

3. Processo Proposto

O processo de extração de relações não-taxonômicas apresentado neste artigo tem seu foco em conceitos da língua portuguesa, e ele constitui-se de 5 etapas distintas (Figura 1): (i) Aquisição dos termos com informações contextuais; (ii) Eliminação de termos com informações faltantes; (iii) Identificação dos conceitos do domínio; (iv) Extração dos candidatos a relações não-taxonômicas; e (v) Visualização das relações extraídas.

Conforme verificou-se através dos trabalhos similares, os verbos são os elementos fundamentais na identificação de relações não-taxonômicas entre conceitos de um domínio. É válido também salientar que na língua portuguesa uma oração tem como estrutura sintática básica "Sujeito + Verbo + Objeto". Constatou-se então que a tabela de termos com informações contextuais produzida pela ferramenta ExATOlp provê todas informações necessárias para a extração de relações não-taxonômicas. Sendo assim, na primeira etapa do processo é feita a aquisição dos termos e todas as suas informações contextuais da lista produzida pelo ExATOlp.

Embora na etapa de aquisição sejam extraídas todas informações contextuais dos termos, para a extração de relações não-taxonômicas são utilizados apenas o termo, sua função sintática (sujeito ou objeto) e o verbo na forma canônica com o qual ele se relaciona. A aquisição de todas informações contextuais é feita com o objetivo de prover informações adicionais aos conceitos que se relacionam para o desenvolvimento de futuras aplicações linguísticas.

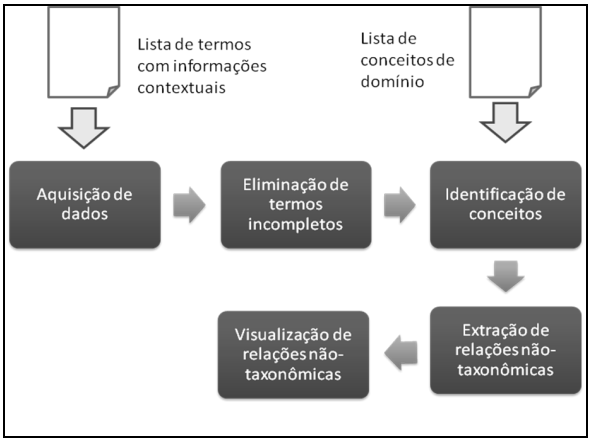

Figura 1. Processo de extração de relações não-taxonômicas

Na segunda etapa são eliminados todos os termos que não possuem verbos associados, ou que o processo de extração não conseguiu identificar a função gramatical. Esta é uma etapa de importante para a rapidez de execução do processo, pois evita que trabalhe-se com termos incompletos. Como o objetivo do processo é extrair relações não-taxonômicas entre conceitos, é necessário processar apenas os termos que foram considerados conceitos de um domínio. Para isto, na terceira etapa do processo os termos são comparados aos conceitos extraídos pelo processo proposto por Lopes et al. (2012) e elimina-se todos os termos que não são classificados como conceitos.

Na etapa de extração de relações não-taxonômicas, são identificados os conceitos que se relacionam através de um mesmo verbo. Estes conceitos são então classificados pela sua função gramatical como sujeitos ou objetos. São definidas então triplas no formato < *Conceito Sujeito, Verbo, Conceito Objeto* > produzidas pelo produto cartesiano entre os conceitos considerados sujeitos e objetos e o verbo que os relaciona. As triplas definidas nesta etapa representam candidatos a relações não-taxonômicas do domínio. Para que seja possível explorar as relações extraídas pelo processo, a última etapa tem seu foco em permitir que usuários (especialistas) visualizem-nas de forma simplificada.

4. Experimento

Com o objetivo de avaliar o funcionamento do processo proposto, foi desenvolvido um sistema computacional para operacionalizar cada uma de suas etapas. Através do sistema desenvolvido, foi realizado um experimento com uma lista de termos, conceitos e informações contextuais produzida pelo ExATOlp a partir de um *corpus* de Geologia. Este *corpus* contém 234 textos, 69.461 frases e 2.010.527 palavras.

Na execução da etapa de aquisição de dados foram encontradas 255.816 ocorrências de termos com informações contextuais. Com o objetivo de eliminar os

termos com informações faltantes, foi executada a segunda etapa do processo, restando 68.831 ocorrências. Na terceira etapa, os termos foram comparados com os conceitos do domínio identificados pelo ExATOlp. Todos os termos que não constavam nessa lista de conceitos foram eliminados, restando então 18.025 ocorrências de conceitos com suas informações contextuais. Sobre estas ocorrências foi realizada a quarta etapa do processo, que produziu 270.197 triplas candidatas a instâncias de relações não-taxonômicas do domínio de Geologia que correspondem a 418 relações distintas.

A quinta e última etapa não foi experimentada, porém para que fosse possível visualizar as triplas candidatas foi desenvolvida uma aplicação Web a ser utilizada pelos especialistas do domínio. Essa aplicação apresenta os dados na forma de um dicionário, permitindo a exploração das instâncias das relações. A Figura 2 apresenta um exemplo onde visualiza-se o conceito "evento vulcânico" (sujeito) e as instâncias (cada objeto) da relação "provocar" (verbo), ou seja, tudo que "eventos vulcânicos" pode "provocar".

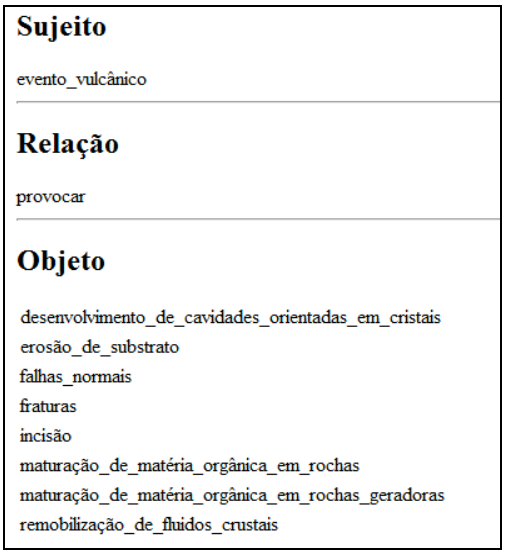

Figura 2. Visualização das relações não-taxonômicas extraídas

8. Conclusões

Conforme pode ser visto através da execução do experimento com o *corpus* de Geologia, o processo proposto permite a extração de relações não-taxonômicas tendo por base a lista de conceitos e suas informações contextuais gerada pelo ExATOlp, ou seja, as informações contextuais dos conceitos disponibilizadas pelo ExATOlp provêem dados suficientes para a descoberta de relações não-taxonômicas.

Um trabalho futuro será avaliar a relevância das relações extraídas para o domínio. Dessa forma, assim como em trabalhos similares, é necessário uma etapa de avaliação por especialistas do domínio das relações extraídas. Portanto, as 418 relações extraídas no processo proposto serão consideradas candidatas e somente após esta etapa de avaliação serão consideradas relações não-taxonômicas do domínio. Outra possibilidade é o uso de pontos de corte, a exemplo do que já foi desenvolvido para selecionar conceitos [Lopes et al. 2010].

Outro trabalho futuro planejado é aplicação do processo proposto para outros *corpora*, como os utilizados por Lopes (2012) relativos às áreas de Pediatria, Modelagem Estocástica, Mineração de Dados e Processamento Paralelo.

Referências

Biemann, C. (2005) *Ontology learning from text: a survey of methods*. LDV Forum, 20: 75-96.

Brewster, C.; Ciravegna, F.; Wilks, Y. (2003) *Background and foreground knowledge in dynamic ontology construction*. Proc. of the SIGIR Semantic Web Workshop.

Chung, T. M. (2003) *A corpus comparison approach for terminology extraction*. Terminology, 9: 221-246.

Cimiano, P.; Volker, J.; Studer, R. (2006) *Ontologies on demand? - a description of the state-of-the-art, applications, challenges and trends for ontology learning from text*. Information, Wissenschaft und Praxis, 57: 315-320.

Gruber T. (1993) *Toward principles for the design of ontologies used for knowledge sharing*. International Journal Human-Computer Studies, 43: 907-928.

Guarino, N. e Welty, C. (2002) *Evaluating ontological decisions with OntoClean*. Communications of the ACM, 45(2): 61-65.

Kavalec M.; Maedche, A.; Svátek, V. (2004) *Discovery of lexical entries for non-taxonomic relations in ontology learning*. Proc. of Int. Conf. on Current Trends in Theory and Practice of Computer Science (SOFSEM), LNCS 2932: 249-256.

Lopes, L.; Vieira, R.; Finatto, M. J.; Martins, D. (2010) *Extracting compound terms from domain corpora*. Journal of the Brazilian Computer Society, 16(4): 247-259.

Lopes, L. (2012) Extração automática de conceitos a partir de textos na língua portuguesa. 156 f. Tese de doutorado em Ciência da Computação - FACIN, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

Lopes, L.; Fernandes, P.; Vieira, R. (2012) Domain term relevance through tf-dcf. Proc. of Int. Conf. on Artificial Intelligence (ICAI).

Maedche A. e Staab S. (2000) *Mining non-taxonomic conceptual relations from text*. Proc. of the 12th European Knowledge Acquisition Workshop, Juan-les-Pins.

Pantel, P. e Lin, D. (2001) *A statistical corpus-based term extractor*. Proc. of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence, ACM Press: 36-46.

Sanchéz, D. e Moreno, A. (2008) *Learning non-taxonomic relationships from web documents for domain ontology construction*. Data & Knowledge Eng, 64: 600-623.

Serra, I. e Girardi, R. (2011) *A process for extracting non-taxonomic relationshop of ontologies from text*. Intelligent Information Management, 3: 119-124.

Schutz, A. e Buitelaar, P. (2005) RelExt: *A tool for relation extraction in ontology extension*. Proc. of the Fourth Int. Semantic Web Conference: 593-606.

Villaverde, J.; Persson, A.; Godoy, D.; Amandi, A. (2009) *Supporting the discovery and labeling of non-taxonomic relationships in ontology learning*. Experts Systems with Applications, 36: 10288-10294.

Weichselbraun, A.; Wohlgenannt, G.; Scharl, A.; Granitzer, M.; Neidhart, T.; Juffinger, A. (2009) *Discovery and evaluation of non-taxonomic relations in domain ontologies*. Int. Journal of Metadata, Semantics and Ontologies, 4(3): 212-222.