LPhDic – Ferramenta para Extração de Conceitos a partir de Textos

Helen de Cássia S. da Costa, Ivan R. Guilherme

Departamento de Estatística, Matemática Aplicada e Computação, IGCE, UNESP – Rio Claro, SP – Brasil

helen.c.s.costa@gmail.com, ivan@rc.unesp.br

Abstract. *This paper presents the LPhDic tool, used for extraction of terms and concepts, the initial phase of the acquisition of ontologies from texts. In particular, LPhDic was developed to improve the methods of linguistic annotation and extraction of concepts from an existing tool, PhDic, used in the construction of knowledge and ontologies from technical reports of drilling and oil production.*

Resumo. *Este artigo apresenta a ferramenta LPhDic, utilizada para extração de termos e conceitos, fase inicial do processo de aquisição de ontologias a partir de textos. Em particular, LPhDic foi desenvolvida para aprimoramento dos métodos de anotação linguística e extração de conceitos de uma ferramenta já existente, PhDic, utilizada na construção do conhecimento e ontologias a partir de relatórios técnicos de perfuração e produção de petróleo.*

1. Introdução

Na busca de suprir a necessidade de estruturas mais consistentes de representação do conhecimento disponível na Internet, a Web Semântica propõe novas ferramentas para extrair e representar o conhecimento expresso em textos através do uso de ontologias. Porém, a construção de ontologias é um processo custoso, tornando essencial o uso de ferramentas que possam auxiliar no desenvolvimento das atividades deste processo.

Neste sentido, este artigo apresenta a ferramenta LPhDic (*Linguistic PhDic*), utilizada para extração de conceitos, fase inicial do processo de aquisição de ontologias a partir de textos. Em particular, LPhDic foi desenvolvida para aprimoramento dos métodos de anotação linguística e extração de conceitos de uma ferramenta já existente, PhDic (*Phrase Dictionary*) [Guilherme *et al.* 2006], utilizada na construção do conhecimento e ontologias a partir de relatórios técnicos de perfuração e produção de petróleo.

2. Metodologia

Esta nova abordagem para extração de conceitos utiliza uma metodologia baseada em informações linguísticas, combinada com medidas estatísticas.

O primeiro ponto a ser considerado é o formato dos textos de entrada utilizados na aplicação. Optou-se por adotar um formato em que os textos já estivessem anotados com informações linguísticas. Por ter sido a ferramenta mais utilizada nos trabalhos estudados, o pré-processamento dos relatórios técnicos foi feito através do analisador

sintático PALAVRAS [Bick 2006] e o formato de saída utilizado como entrada para a aplicação foi o Tiger-XML [Konig *et al.* 2003].

A primeira fase do processo é a extração de termos, responsável pela extração de termos simples (palavras) dos documentos, considerando as classes gramaticais que eles pertencem. São considerados termos relevantes somente palavras que pertençam às classes gramaticais que normalmente representam algum tipo de conceito, como substantivo, adjetivo e advérbio. Algumas restrições também são consideradas na extração de termos: a forma canônica, o tamanho e o tipo da palavra. Estas restrições são usadas no OntoLP [Ribeiro Junior 2008] e por gerar bons resultados foram adotadas neste trabalho. Após as restrições, para cálculo de relevância de termos são aplicados os métodos Frequência Relativa (FR) e TFIDF (*Term Frequence-Inverse Document Frequence*). Em seguida, os termos são reorganizados em ordem decrescente de relevância e disponibilizados para o usuário que pode excluir da lista apresentada termos que considerar desnecessários ou incorretos. Esta listagem final é utilizada como entrada para a próxima fase.

A fase seguinte é chamada extração de conceitos, responsável pela extração de termos compostos ou multi-palavras, baseada em padrões morfossintáticos presentes nas sentenças. Os padrões adotados são as regras propostas por Baségio [Baségio 2006]. Porém, são extraídos somente termos compostos que tenham pelo menos uma palavra pertencente à lista de termos gerada na etapa anterior. Nesta fase também são consideradas as restrições feitas na extração de termos. Após este processo, também são aplicadas as medidas estatísticas FR e TFIDF para avaliação da relevância de conceitos extraídos. Por fim, a lista gerada é reorganizada em ordem decrescente de relevância, e em seguida, o usuário também pode excluir conceitos que considerar desnecessários ou irrelevantes para o domínio.

Os resultados finais do processo de extração de termos e conceitos são disponibilizadas em arquivos de texto, que contém o conceito gerado seguido de sua frequência relativa e TFIDF.

3. Análise de Resultados

Para análise de resultados foram processados 2.088 textos referentes a relatórios técnicos de anormalidades na perfuração e produção de petróleo. O número de termos e conceitos geradas pela ferramenta LPhDic, sem a intervenção do usuário, são apresentadas na Tabela 1.

Tabela 1. Número de conceitos extraídos pelo LPhDic

Lista	Nro. de Extraídos
unigramas	728
bigramas	301
trigramas	1.245
quadrigramas	249
pentigramas	7
hexigramas	0

O desempenho da ferramenta LPhDic foi avaliado através de comparações entre as listas geradas de unigramas e bigramas com as respectivas listas de referência extraídas manualmente por um engenheiro de petróleo, especialista do domínio. Além

disto, também foi feita uma comparação do desempenho com as ferramentas PhDic e ExATOlp [Lopes *et al.,* 2009]. Para fazer as comparações foram usadas três medidas estatísticas: Precisão, Abrangência e *F-mesuare*. Os resultados são apresentados na Tabela 2.

Tabela 2. Comparação de Resultados

LPhDic						
Tipo de Lista	**Termos Extraídos**	**Termos Corretos**	**Lista de Referência**	**Precisão (%)**	**Abrangência (%)**	***F-measure* (%)**
unigramas	728	204	654	28,02	31,19	29,52
bigramas	301	235	528	**78,07**	44,5	56,69
PhDic						
unigramas	947	322	654	34	**49,23**	**40,22**
bigramas	2383	198	528	8,3	37,5	13,59
ExATOlp						
unigramas	464	175	654	**37,71**	26,76	31,3
bigramas	463	281	528	60,69	**53,22**	**56,71**

Para a extração unigramas, a precisão da LPhDic foi menor do que a precisão das outras ferramentas e a abrangência foi maior do que a abrangência da ExATOlp, sendo a média harmônica entre essas duas medidas menor do que a média das outras duas ferramentas. Já para bigramas, a precisão da LPhDic foi maior do que a precisão das outras ferramentas e a abrangência foi maior do que a abrangência do PhDic, sendo a média harmônica entre essas duas medidas praticamente igual a da ExATOlp, que obteve melhor média.

4. Conclusão

A partir da comparação feita entre a ferramenta LPhDic e a PhDic, por meio da análise de resultados, foi possível observar que, apesar do PhDic ter obtido um resultado satisfatório para extração de unigramas, com uma abrangência de aproximadamente 50%, a ferramenta LPhDic obteve melhores resultados para extração de bigramas, com precisão de aproximadamente 78%, contra 8,3% do PhDic. Além disto, através da adoção do anotador sintático PALAVRAS, foi possível melhorar a automatização do processo de extração de conceitos, haja vista que no processo usado pelo PhDic os termos simples extraídos são anotados manualmente com uma sintaxe pré definida pelo usuário. Porém, é importante ressaltar que a dependência de um analisador sintático pago pode ser considerada como uma limitação da ferramenta LPhDic.

Referências

Baségio, T. L. (2006) “Uma Abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil”, Dissertação (Mestrado). Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS.

Bick, E. (2000) “The parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework”, PhD thesis, Arhus University.

Guilherme, I. R., Serapiao, A. B. S., Rabelo, C. e Mendes, J. R. P. (2006) “An Ontology Based for Drilling Report Classification”, Em: Lecture Notes in Computer Science, v. 1, p.inicial 1037, p.final 1046, ISSN: 0302-9743.

Konig, E., Lezius, W. e Voormann, H. (2003) “TIGERSearch 2.1 – User’s Manual”, IMS, University of Stuttgart, http://www.ims.unistuttgart.de/projekte/TIGER/TIGERSearch/doc/html/. Acesso em: 27 de Set. 2009.

Lopes, L., Fernandes, P., Vieira, R. e Fedrizzi, G. (2009) “ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora”, Em: LTC'09 - 4th Language and Technology Conference, 2009, Poznan, 2009, Poznan. Proceedings of the Fourth Language and Technology Conference. Poznan : Adam Mickiewicz University, 2009. p. 427-431.

Ribeiro Junior, L. C. (2008) “OntoLP: Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa”, 68 f. Dissertação (Mestrado) - Universidade Do Vale Do Rio Dos Sinos, São Leopoldo, 2008, http://www.inf.pucrs.br/~ontolp/downloads-ontolpplugin.php. Acesso em: 31 Jul. 2009.