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

Geração automática de glossários de termos específicos de um *corpus* de Geologia

Igor da Silveira Wendt (PUCRS) igor.wendt@acad.pucrs.br

Lucelene Lopes (PUCRS) lucelene.lopes@pucrs.br

Daniel Martins (PUCRS) daniel.martins@acad.pucrs.br

Renata Vieira (PUCRS) renata.vieira@pucrs.br

Vera Lúcia Strube de Lima vera.strube@pucrs.br

Resumo: Este artigo descreve um trabalho focado na construção automática de um glossário de domínio específico. Para desenvolver esse trabalho foi utilizado um corpus de geologia geral, de onde foram extraídos termos relevantes para posterior recuperação de definições desses termos através da web. Ainda que este experimento tenha um foco na construção de glossários, os termos extraídos e suas definições podem auxiliar em outras tarefas, como por exemplo, na construção de ontologias, bem como no desenvolvimento de material de auxilio à tradução.

1. Introdução

Atualmente diversos grupos de pesquisas têm características multidisciplinares. Para o desenvolvimento desse tipo de pesquisa é necessário que pesquisadores oriundos de diversas áreas de atuação interajam entre si. Essa interação é dificultada pelo conjunto de conceitos que cada pesquisador traz da sua área.

Neste sentido, possuir um glossário de termos específicos permite uma melhor comunicação entre as diversas áreas envolvidas. Este artigo apresenta um experimento focado na extração de termos de um *corpus* em língua portuguesa para a geração automática de um glossário no domínio da geologia.

Os termos candidatos chamados *definiendum* (latim: termo a ser definido) (MEYER 2001), são extraídos a partir de um *corpus* sintaticamente anotado no domínio de geologia geral, utilizando-se a ferramenta ExATOlp - Extrator Automático de Termos para Ontologias em Língua Portuguesa (LOPES et.al. 2009). Posteriormente buscou-se a definição, chamada *definiens* (latim: aquilo que caracteriza o que é definido) de cada um desses termos através de informações especializadas disponíveis na Internet. Essas informações são obtidas de fontes com diferentes graus de confiabilidade, indo desde glossários específicos desenvolvidos manualmente por especialistas do domínio, até definições da Wikipédia.

Esse artigo está organizado na seguinte maneira: a Seção 2 apresenta alguns trabalhos relacionados ao desenvolvido neste artigo; a Seção 3 apresenta detalhes sobre o processo de construção automática dos glossários; por fim, a Seção 4 conclui apresentando uma avaliação dos resultados obtidos e possíveis trabalhos futuros.

2. Trabalhos Relacionados

No trabalho proposto em (SILVA 2008), uma ferramenta foi desenvolvida para auxiliar engenheiros de ontologia. Esta ferramenta é utilizada como um *Plug-in* para o Protégé[1], e tem como foco a obtenção automática de definição de conceitos presentes em uma Ontologia. A ferramenta consiste em utilizar os conceitos de uma ontologia como termos para realizar uma busca por suas definições na Web. Para recuperar essas definições são utilizados sites como o Wikionário, Wikipédia e Google. No entanto, este trabalho difere do aqui apresentado em dois pontos. Inicialmente, assume-se como conhecido os conceitos a definir, enquanto no nosso trabalho realiza-se uma etapa de extração e seleção de termos. Igualmente, nosso trabalho difere de (SILVA 2008) por utilizar fontes de definições com diferentes graus de confiabilidade.

O trabalho proposto em (SCLANO e VELARDI 2007) apresenta uma ferramenta, TermExtract, para a extração de definições através da Internet. O processo utilizado é semelhante ao nosso, pois um corpus é construído sobre o domínio de interesse e termos candidatos são extraídos segundo um processo híbrido que leva em consideração informações linguísticas e decide os melhores candidatos a partir de critérios estatísticos. Uma diferença é que estes termos candidatos são submetidos a buscas genéricas na Internet procurando estruturas frasais que contenham o termo (*T*) seguido de verbos que sugiram definições, como por exemplo: “*T* is a”, “*T* is an”, “*T* are the”, “*T* defines”, “*T* refers to”, “*T* concerns”, “*T* is the” e “*T* is any”. Posteriormente, as definições recuperadas passam por dois filtros, um chamado de *Stylistic filter* onde se procura encontrar definições bem formadas, o outro filtro, chamado de *Domain filter*, visa remover as definições que não são pertinentes ao domínio utilizado.

Em (PARK *et al.* 2002) apresenta-se um trabalho bastante semelhante à nossa proposta de construção de glossários, pois ambos trabalhos descrevem métodos baseados em textos específicos do domínio do qual se pretende extrair os conceitos. No entanto, em (PARK et al. 2002) os termos candidatos a *definiendum* passam por um processo de

agregação de termos que são apenas variantes de um mesmo conceito, como abreviações ou erros de digitação. Neste sentido, os termos brutos são ranqueados através de métricas como grau de coesão do termo, para determinar variantes, e grau de especificidade, para escolher os termos mais relevantes. No trabalho que propomos, a análise de variantes limita-se a localização de termos que possuem a mesma forma canônica. Outra diferença entre nossa proposta e o trabalho de (PARK et al. 2002) é que o nosso processo é completamente automático, não sendo necessária a revisão de especialista do domínio.

3. Processo de Construção Automática de Glossários

O processo utilizado se divide em três passos que são descritos a seguir (Figura 1). O processo inicia com uma extração de termos candidatos a partir de um *corpus* de domínio (Passo 1). Em seguida uma base de definições é construída a partir de dois

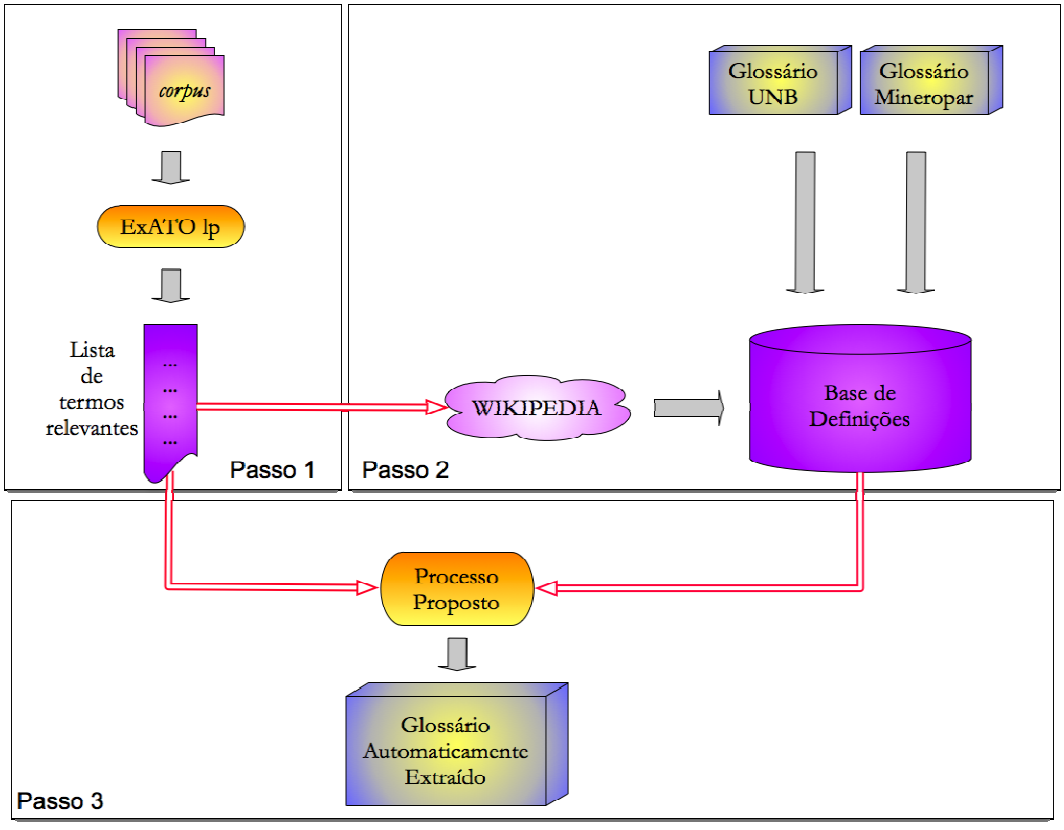

Figura 1. Processo de Construção Automática Utilizado.

[1] http://protege.stanford.edu/

glossários existentes e de definições encontradas na Wikipédia para os termos extraídos (Passo 2). Finalmente, procura-se entre as definições existentes na base a melhor definição, segundo critérios, para cada um dos termos extraídos (Passo 3). Cada um destes passos é detalhado nas seções a seguir.

3.1. Passo 1 - Extração de Termos Relevantes do *corpus*

O primeiro passo do processo de extração consiste em extrair de um *corpus* anotado uma lista de termos relevantes. Este passo é a base para a geração de glossários, pois o resultado das próximas etapas é altamente dependente dos resultados aqui obtidos. A meta principal dessa etapa é a extração de termos candidatos a obtenção de definições. Neste sentido, esta seção descreve o *corpus* e a ferramenta de extração utilizada.

O *corpus* utilizado nos experimentos é composto por 140 textos em português da área de geologia geral em um total de aproximadamente um milhão de palavras. Esse *corpus* foi desenvolvido dentro de um trabalho de doutorado que está inserido no grupo de Processamento de Linguagem Natural da PUCRS.

Para construção do *corpus* alguns critérios foram adotados para que fossem coletados apenas textos científicos (artigos, teses e dissertações). Os textos encontrados, geralmente arquivos *.pdf* foram posteriormente encaminhados para especialistas do domínio que avaliaram e aprovaram esses textos sob o ponto de vista de serem adequados e possuir qualidade dentro da área específica. Após a análise dos especialistas os documentos aprovados foram convertidos para arquivos no formato *.txt*, com o auxilio da ferramenta Entrelinhas (SILVEIRA 2008). Com o propósito de se trabalhar com textos anotados com informação sintática das frases, retirou-se título de seção, referências bibliográficas, tabelas, gráficos e algumas seções irrelevantes para o domínio como dedicatória, agradecimentos, sumários, etc. Na Tabela 1 são apresentados os detalhes sobre a composição do *corpus*.

Tabela 1: Composição do *Corpus.*

Tipo	Número de textos	Número de tokens
Artigo	119	815.381
Tese	11	110.788
Dissertação	10	88.528

Após todo o processo descrito no parágrafo anterior, ou seja, com o *corpus* limpo, foi utilizado o *parser* PALAVRAS (ECKHARD 2000) para o processo de anotação linguística. O *parser* gera como saída um documento em formato XML (*eXtensible Markup Language*). Este documento XML é estruturado em formato de árvore e em cada frase do documento analisado são acrescentadas informações morfológicas, sintáticas e semânticas sobre palavras e as frases dos textos.

Uma vez o *corpus* definido e anotado linguisticamente, a extração de termos de fato é feita através da ferramenta ExATOlp. Essa ferramenta foi desenvolvida para o processamento de textos em português, ela recebe um *corpus* anotado e extrai automaticamente todos os sintagmas nominais (SN).

A Figura 2 apresenta uma frase anotada pelo PALAVRAS. A frase é decomposta em tokens entre as tags “<terminals>” e “</terminals>”, cada token carrega junto consigo informações morfológicas. Entre as tags “<nonterminals>” e “</nonterminals>” são apresentados os SN. Os SN são identificados através do atributo cat=‘NP’ da tag “<nt ...>”. Em seguida é indicado através do id das tokens, quais os tokens que formam o SN, por exemplo, a tag <nt id=‘s6_503’ cat=‘NP’> indica que os tokens de id=‘s6_01’, id=‘s6_02’ e id=‘s6_03’ seguidos de id=s6_4 e id=s6_6 formam um SN.

Baseado em informações linguísticas, o ExATOlp utiliza algumas heurísticas para refinar o processo de extração de termos. São apresentadas algumas destas heurísticas abaixo, mais detalhes são apresentados em (LOPES et al. 2010).

- São eliminados termosextraídos como SN, mas que terminam com preposição. e.g., “rocha acrescida de”, “dosagem diária para”;
- São eliminados SN que possuem números, e.g., “década de 50”, “dois estudos”;
- São excluídos os SN cujo núcleo não for substantivo, nem nome próprio, nem adjetivo, e.g., “valor superestimado”, “observado por outros”;
- SN que começam com artigos são armazenados sem a primeira palavra (o artigo), e.g., “a rocha magmática” é armazenado apenas como “rocha magmática”.

```
<s id="s6" ref="6" source="Running text" forest="1" text="O elemento lamoso FF (finos da planície de inundação)
                                                          ocorre em toda área estudada.">
    <graph root="s6_500">
        <terminals>
            <t id="s6_1" word="O" lemma="o" pos="art" morph="M S" sem="--" extra="--"/>
            <t id="s6_2" word="elemento" lemma="elemento" pos="n" morph="M S" sem="ac" extra="--"/>
            <t id="s6_3" word="lamoso" lemma="lamoso" pos="adj" morph="M S" sem="--" extra="DERS np-close"/>
            <t id="s6_4" word="FF" lemma="FF" pos="prop" morph="M S" sem="--" extra="org np-long"/>
            <t id="s6_5" word="(" lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
            <t id="s6_6" word="finos" lemma="fino" pos="adj" morph="M P" sem="--" extra="np-close"/>
            <t id="s6_7" word="de" lemma="de" pos="prp" morph="--" sem="--" extra="sam-"/>
            <t id="s6_8" word="a" lemma="o" pos="art" morph="F S" sem="--" extra=" -sam"/>
            <t id="s6_9" word="planície" lemma="planície" pos="n" morph="F S" sem="Ltop" extra="--"/>
            <t id="s6_10" word="de" lemma="de" pos="prp" morph="--" sem="--" extra="np-close"/>
            <t id="s6_11" word="inundação" lemma="inundação" pos="n" morph="F S" sem="event" extra="--"/>
            <t id="s6_12" word=")" lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
            <t id="s6_13" word="ocorre" lemma="ocorrer" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="fmc mv"/>
            <t id="s6_14" word="em" lemma="em" pos="prp" morph="--" sem="--" extra="--"/>
            <t id="s6_15" word="toda" lemma="todo" pos="pron-indef" morph="DET F S" sem="--" extra="quant"/>
            <t id="s6_16" word="área" lemma="área" pos="n" morph="F S" sem="L" extra="--"/>
            <t id="s6_17" word="estudada" lemma="estudado" pos="adj" morph="F S" sem="--" extra="np-close"/>
            <t id="s6_18" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
        </terminals>

        <nonterminals>
            <nt id="s6_500" cat="s">
                <edge label="UTT" idref="s6_501"/>
            </nt>
            <nt id="s6_501" cat="x">
                <edge label="fA" idref="s6_505"/>
                <edge label="H" idref="s6_7"/>
                <edge label="DP" idref="s6_506"/>
            </nt>
            <nt id="s6_502" cat="fcl">
                <edge label="S" idref="s6_503"/>
            </nt>
            <nt id="s6_503" cat="np">
                <edge label="DN" idref="s6_1"/>
                <edge label="H" idref="s6_2"/>
                <edge label="DN" idref="s6_3"/>
                <edge label="DN" idref="s6_504"/>
            </nt>
            <nt id="s6_504" cat="np">
                <edge label="H" idref="s6_4"/>
                <edge label="DNc" idref="s6_6"/>
            </nt>
            <nt id="s6_505" cat="pp">
            </nt>
            <nt id="s6_506" cat="np">
                <edge label="DN" idref="s6_8"/>
                <edge label="H" idref="s6_9"/>
                <edge label="DN" idref="s6_507"/>
            </nt>
            <nt id="s6_507" cat="pp">
                <edge label="H" idref="s6_10"/>
                <edge label="DP" idref="s6_11"/>
                <edge label="P" idref="s6_13"/>
                <edge label="fA" idref="s6_508"/>
            </nt>
            <nt id="s6_508" cat="pp">
                <edge label="H" idref="s6_14"/>
                <edge label="DP" idref="s6_509"/>
            </nt>
            <nt id="s6_509" cat="np">
                <edge label="DN" idref="s6_15"/>
                <edge label="H" idref="s6_16"/>
                <edge label="DN" idref="s6_17"/>
            </nt>
        </nonterminals>
    </graph>
</s>
```

FIGURA 2 - Exemplo de anotação feita pelo PALAVRAS.

Do processo de extração de termos, sobre o corpus de geologia utilizado neste artigo, foi gerada uma lista contendo um total de 4626 termos, sendo destes, 1673 unigramas (1tokens), 2003 bigramas (2 tokens) e 950 trigramas (3 tokens). Termos com quatro ou mais palavras (tokens) não foram considerados.

3.2. Passo 2 - Construção da Base de Definições

Neste passo de construção da base de definições foram utilizados dois glossários específicos do domínio disponíveis publicamente na Internet. O primeiro glossário foi desenvolvido por professores do Instituto de Geociências da Universidade de Brasília – UnB e possui 1447 termos com suas definições e ilustrações. O segundo glossário foi desenvolvido pela compania de Mineração do estado do Paraná – MINEROPAR[2] e possui 3078 termos e suas definições.

Como nem todos os termos da lista extraída do corpus foram encontrados nesses glossários, a base de definições foi complementada através de uma busca específica desses termos na Wikipédia. Dessa forma é possível encontrar diversas definições para o mesmo termo, logo o terceiro e último passo consiste em escolher a definição mais adequada para os termos extraídos como é visto na seção a seguir.

3.3. Passo 3 - Escolha das Melhores Definições

Inicialmente os dois glossários específicos são considerados corretos, logo para cada termos da lista extraída, quando existe uma definição em um dos glossários, as definições da Wikipédia não são consideradas. Igualmente, o glossário da UnB, por ser um trabalho acadêmico, é considerado mais adequado. Por isso suas definições são consideradas prioritárias às do glossário da MINEROPAR.

Na existência de definições somente através da Wikipédia se faz um processo de escolha entre as definições disponíveis através de um cálculo do índice de pertinência ao domínio. Este índice é calculado como o número de termos extraídos do *corpus* presentes no texto de cada uma das definições, dividido pelo total de palavras do texto da definição. Desta forma, escolhe-se como a melhor definição aquela onde outros termos também extraídos do *corpus* são mais frequentes. A divisão pelo número de palavras no texto de cada definição visa não beneficiar definições muito extensas.

Finalmente, caso um termo extraído não possua nenhuma definição encontrada, nem nos glossários, nem na pesquisa sobre a Wikipédia, este termo é descartado, ou seja, ele não é incluído no glossário final.

[2] http://www.mineropar.pr.gov.br

4. Conclusão

A aplicação do processo de extração ao corpus resultou em 4626 termos de até três palavras, sendo 1673 unigramas, 2003 bigramas e 950 trigramas. Termos com 4 ou mais palavras foram desconsiderados neste experimento de geração de glossários.

Considerando os dois glossários (UnB e MINEROPAR) como uma lista de referencia, temos 2771 unigramas, 1141 bigramas e 450 trigramas

A comparação das listas extraídas com os termos presentes em qualquer um dos dois glossários resultou em que 926 unigramas, 458 bigramas e 142 trigramas. Considerando estes glossários como listas de referência, este resultado indica uma precisão de 55% para a extração de unigramas, 23% para bigramas e 15% para trigramas. Estes valores de precisão indicam um grande número de termos compostos (bigramas e trigramas) extraídos ausentes dos glossários. O que é razoável dado o baixo número de termos compostos nos glossários. Por outro lado a abrangência é mais uniforme ficando em 33% para unigramas, 40% para bigramas e 32% para trigramas.

Procurando definições dos termos extraídos não encontrados nos glossários na Wikipedia, os números de unigramas, bigramas e trigramas para os quais se obteve uma definição subiu para 1367, 488 e 268 termos, respectivamente. Com estes novos valores, a precisão dos termos extraídos sobe agora para 82% para unigramas, 24% bigramas e 28% para trigramas. Este aumento de precisão foi bastante grande para unigramas e trigramas.

Tabela 2: Composição do *Corpus.*

	Termos extraídos	Termos de referência	Intersecção	Abrangência	Precisão
Unigramas	1673	2771	926	33%	55%
Bigramas	2003	1141	458	40%	23%
Trigramas	950	450	142	32%	15%
Total	4626				

Observando os termos extraídos que não estavam presentes nos dois glossários de referência se observa que para 2503 deles também não foi encontrada nenhuma definição na Wikipédia. Este número é relativamente alto, pois um pouco mais da

metade dos termos extraídos não foi encontrada nem nos glossários especializados, nem na Wikipédia. Percebe-se que em particular termos compostos foram consideravelmente menos frequentes nos glossários e Wikipédia do que nas listas de termos extraídos.

No entanto, este resultado não é necessariamente um problema do processo proposto. Não encontrar definições pode indicar que o processo de extração automática de termos a partir de corpus especializadoé mais abrangente que glossários e páginas da Wikipédia atualmente disponíveis.

Cabe salientar que as listas de termos extraídos diz respeito a apenas um conjunto finito de textos sobre o domínio, ou seja, os termos relevantes para a totalidade do domínio tedem a ser ainda maiores do que o que foi obtido pela análise do corpus utilizado.

A principal contribuição deste trabalho situa-se na proposta de um processo automático de construção de um glossário a partir de fontes existentes e um *corpus* do domínio. Este processo pode ser mais refinado com uma análise linguística dos textos selecionados como definições, mas o processo atual já contribui com a análise de um índice de pertinência ao domínio baseado nos termos extraídos automaticamente.

Estes resultados formam a base para um estudo mais aprofundado que poderá integrar outras formas de busca de definições que permitirão uma melhor qualidade dos glossários criados automaticamente. Apesar disto, os resultados obtidos para este *corpus* de Geologia apresentaram uma boa qualidade aumentando bastante o número de unigramas (de 55% para 82%) e trigramas (de 15% para 28%) com definições em comparação com os glossários manuais utilizados como referência.

Referências

ECKHARD, B. **The Parsing System" Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Aarhus University Press, 2000

LOPES, L., Fernandes, P., Vieira, R., and Fedrizzi, G. **Exato lp – an automatic tool for term extraction from portuguese language corpora**. In Proceedings of the Fourth Language and Technology Conference, volume 1, pages 1–5. LTC'09, 2009a.

LOPES, L., Vieira, R., Finatto, M. J., Zanette, A., Martins, D., and Ribeiro Jr, L. C. **Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area**. RECIIS, 3(1):72–84, 2009b.

MEYER, I. **Extracting knowledge-rich contexts for terminography. Recent advances in computational terminology**, pages 279–302, 2001.

PARK, Y.; BIRD, R.; BOUGAREV, B. Automatic Glossary Extraction: Beyond Terminology Identification. Proceedingings of the 19th COLING, Taipei, Taiwan, 2002.

SILVA, M. **Extração automática de descrição de conceitos de ontologias em textos de língua portuguesa**. Monografia, Universidade do Vale do Rio dos Sinos, 2008.

SILVEIRA, F. P. **Entrelinhas - uma ferramenta para processamento e análise de corpus**. Dissertação de Mestrado, PUCRS, 2008.

SCLANO, F.; VELARDI, P. **TermExtractor: a web application to learn the common terminology of interest groups research communities**. 9th Conference on Terminology and Artificial Intelligence, TIA 2007, Sophia-Antipolis, 2007.