EXTRAÇÃO AUTOMÁTICA DE SINTAGMAS NOMINAIS PARA CONSTRUÇÃO ONTOLOGIAS

Lucelene Lopes PG (PUCRS)

Renata Vieira (PUCRS)

ExATOlp - Extrator Automático de Termos para Ontologias em Língua Portuguesa é uma ferramenta que recebe um corpus anotado e extrai automaticamente todos os sintagmas nominais (SN) deste texto. A função primária da ferramenta é extrair termos candidatos a conceitos, auxiliando na construção de ontologias, glossários e outros recursos semânticos.

Os sintagmas extraídos são salvos em listas que podem conter tanto os SN na sua forma original no texto, como em sua forma canônica. A ferramenta ainda oferece algumas opções de manipulação usuais para listas de termos como a aplicação de pontos de corte, comparação de listas e cálculo de medidas usuais de precisão e abrangência.

As funcionalidades da ferramenta vão desde tarefas fortemente baseadas em conceitos linguísticos como a extração de sintagmas nominais, até tarefas puramente estatísticas como o cálculo de métricas de avaliação, passando por tarefas como a localização de termos extraídos identificando os textos e frases onde eles ocorrem.

A ferramenta utiliza um conjunto de heurísticas opcionais para refinar o processo de extração. Estas heurísticas tem base linguística com o propósito de eliminar ou refinar SN que não sirvam como possíveis conceitos de uma ontologia, especificamente:

- são eliminados SN que possuem números, por exemplo, “20 anos”, “seis meses”;

- são aceitos apenas sintagmas que possuem letras (acentuadas ou não) ou hífen, ou seja, SN que contém caracteres especiais são eliminados, por exemplo, “dupla mãe/neonato”;

- termos identificados como SN que iniciam com pronomes, “estas condições” e “todas as crianças”, são armazenados sem o pronome;

- termos identificados como SN que terminam com conjunções, por exemplo, “baixo peso e” e “leite materno ou” são armazenados sem a conjunção;

-termos identificados como SN que terminam com preposição, por exemplo, “criança acrescida de” e “dosagem diária para” são armazenados sem a preposição;

- termos identificados como SN que contém artigos são armazenados sem estes artigos, “a cicatriz renal” é armazenado apenas como “cicatriz renal”.

Opcionalmente, ainda é possível escolher armazenar apenas alguns SN sendo critérios o número de palavras que o compõem, a sua classe gramatical e a classe sintática do núcleo do SN. Estas opções são:

- é possível selecionar para extrair apenas SN compostos de números específicos de palavras, por exemplo, pode-se escolher extrair apenas sintagmas compostos de uma, duas e três palavras, ou seja, desprezar sintagmas compostos de quatro ou mais palavras;

- é possível extrair somente SN que aparecem como sujeitos, ou somente SN que aparecem como complementos das orações;

- é possível extrair somente SN que possuem como núcleo substantivos próprios, só substantivos comuns, só adjetivos, só verbos no particípio passado, ou qualquer combinação entre estas.

Em geral, a saída do processo de extração gera uma lista de termos muito extensa, a qual inclui termos relevantes, mas também um número grande de termos irrelevantes. É interessante buscar uma forma de reduzir o tamanho das listas, excluindo o mínimo possível de termos relevantes. Estas listas podem ser facilmente submetidas a pontos de corte que levam em consideração a frequência relativa ou absoluta de cada termo. Desta forma, os termos extraídos são organizados segundo sua frequência no *corpus* e um ponto de corte pode ser aplicado.

ExATOlp disponibiliza as seguintes opções de ponto de corte:

- ponto de corte absoluto segundo a frequência relativa, onde um limiar mínimo (um número real entre 0 e 1) deve ser informado;

- ponto de corte absoluto segundo a frequência absoluta, onde um limiar mínimo (um número inteiro superior a 1) deve ser informado;

- ponto de corte absoluto único, onde um número específico de termos (um inteiro) deve ser informado;

- ponto de corte relativo, onde um percentual do número de termos (um valor entre 0% e 100%) deve ser informado.

Outra funcionalidade da ferramenta é a possibilidade de comparar as listas extraídas com listas de referência. Neste caso as listas são denominadas de lista de referência (LR) e lista de extraídos (LE). ExATOlp faz esta comparação podendo retornar qualquer uma das seguintes listas:

- a intersecção entre elas ($LR \cap LE$);

- a união entre elas ($LR \cup LE$);

- os termos de LR ausentes em LE ($LR - (LR \cap LE)$);

- os termos de LE ausentes em LR ($LE - (LR \cap LE)$).

Com intuito de tornar objetiva a comparação de listas, a ferramenta ExATOlp disponibiliza também o cálculo de métricas quantitativas que expressam a precisão (P) e a abrangência (R) de listas comparadas, bem como o equilíbrio entre estes dois índices denominado *f-measure* (F).

Estas métricas são calculadas pelas seguintes fórmulas:

$$P = \frac{|(LR \cup LE)|}{|LR|} \qquad R = \frac{|(LR \cup LE)|}{|LE|} \qquad F = \frac{2 \; P \; R}{P + R}$$

Outra funcionalidade da ferramenta é a localização de termos extraídos no *corpus*. Nesta funcionalidade a ferramenta permite localizar todos os termos extraídos, ou buscar apenas um termo específico. Enquanto a primeira opção tem uma saída mais voltada a um tratamento computacional posterior, a segunda opção oferece uma interface amigável ao usuário onde são mostradas cada uma das frases onde o termo procurado aparece. Desta forma, a ferramenta permite que as várias ocorrências de um determinado sejam observadas em detalhe.

O desenvolvimento da ferramenta ExATOlp se insere em um trabalho de doutorado com o propósito de gerar automaticamente ontologias em língua portuguesa à partir de *corpus*, logo, novos avanços de pesquisa estão sendo incorporados à ferramenta regularmente. Uma lista completa de funcionalidades, documentação, exemplos de utilização e o próprio *download* de versões da ferramenta ExATOlp podem ser encontrados http://www.inf.pucrs.br/~ontolp/exato.php. Encoraja-se o leitor interessado no assunto a visitar regularmente esta página para acompanhar as evoluções da ferramenta que na sua versão atual esta disponível para as seguintes plataformas computacionais: MSWindows (todas versões), Linux e MacOS.

Palavras-Chave: Extração Automática de Termos, Ontologias, Sintagmas Nominais, ExATOlp