UMA ABORDAGEM PARA A OBTENÇÃO DE PALAVRAS-CHAVE EM ARTIGOS DE NATUREZA BIOTECNOLÓGICA

***Kirill Lassounski*[1], *Sahudy M. González*[2], *Annabell D.R. Tamariz*[3]**

[1]UENF/Laboratório de Ciências Matemáticas, lassounski@gmail.com
[2]UFABC/ Centro de Matemática, Computação e Cognição, sahudy_montenegro@yahoo.com
[3]UENF/Laboratório de Ciências Matemáticas, annabell_brasil@yahoo.es

Resumo - O *National Center for Biotechnology Information* (NCBI) provê informações sobre genes e sequências de proteínas, literaturas científicas, estruturas moleculares, dentre outros recursos relacionados à área de biomedicina. No portal do NCBI, existe um banco de dados chamado PubMed que guarda atualmente cerca de 19 milhões de artigos científicos em inglês. A dificuldade dos pesquisadores ao buscar no PubMed reside em obter os artigos que são realmente relevantes. Geralmente, é feita uma busca no site do PubMed e aparece uma grande lista de artigos linear, na qual se pode escolher um artigo e obter informações sobre ele, o que torna o trabalho de busca muito exaustivo e demorado, pois cada artigo deve ser analisado individualmente. Para tornar o processo de busca no portal mais simples, eficiente e eficaz, propõe-se determinar palavras-chave que descrevem os artigos retornados, a partir de uma busca inicial realizada pelo pesquisador. Este trabalho visa criar um algoritmo para a extração automática de palavras-chave em inglês, a partir dos resumos de artigos retornados de uma pesquisa feita ao PubMed. Os resultados inicias do algoritmo são avaliados utilizando as métricas *precision* e *recall* para determinar a proporção de palavras-chave relevantes que foram extraídas.

Palavras-chave: NCBI, biotecnologia, processamento de linguagem natural, recuperação de informação.

Área do Conhecimento: Ciência da computação e informática.

Introdução

O portal NCBI[1] possui em seu banco de dados PubMed cerca de 19 milhões de artigos científicos em inglês, além de sequências de proteínas, genomas e outros. Estes artigos são buscados diariamente por pesquisadores de todas as partes do mundo em busca de informações que irão ajudá-los em seus estudos. Com esta grande quantidade de dados disponível torna-se muito difícil encontrar artigos relevantes para uma determinada busca. A disponibilização de palavras-chave, que descrevam de uma forma geral os assuntos abordados nos artigos retornados em uma pesquisa, podem facilitar e melhorar o processo de busca neste portal. Alguns artigos já possuem palavras-chave selecionadas manualmente pelos autores, mas a maioria não possui. Fazer o trabalho de seleção de palavras-chave manualmente é muito cansativo e tedioso. Este trabalho propõe a criação de um algoritmo para a extração automática de palavras-chave em inglês, a partir dos resumos de artigos retornados de uma pesquisa feita ao PubMed.

Metodologia

[1]http://www.ncbi.nlm.nih.gov/

Existem duas abordagens para a obtenção de termos relevantes em um texto: a extração de palavras-chave baseada nas próprias palavras contidas no texto e a obtenção a partir de vocabulários controlados (WITTEN, 1999). Neste trabalho foi escolhida a primeira abordagem, pois torna a pesquisa mais flexível e independente de vocabulários controlados que podem ficar desatualizados.

A Figura 1 apresenta o esquema geral do algoritmo proposto para a obtenção das palavras-chave, a partir dos resumos e a Figura 2 destaca um exemplo.

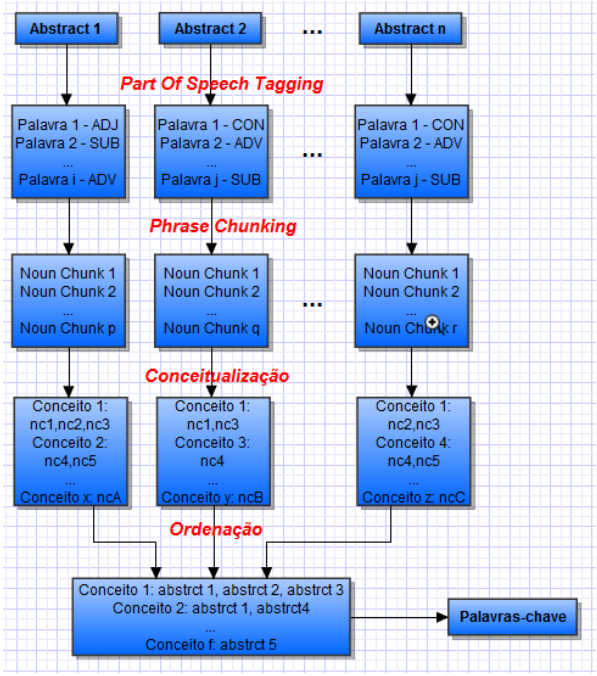

Figura 1- Processo de obtenção de termos.

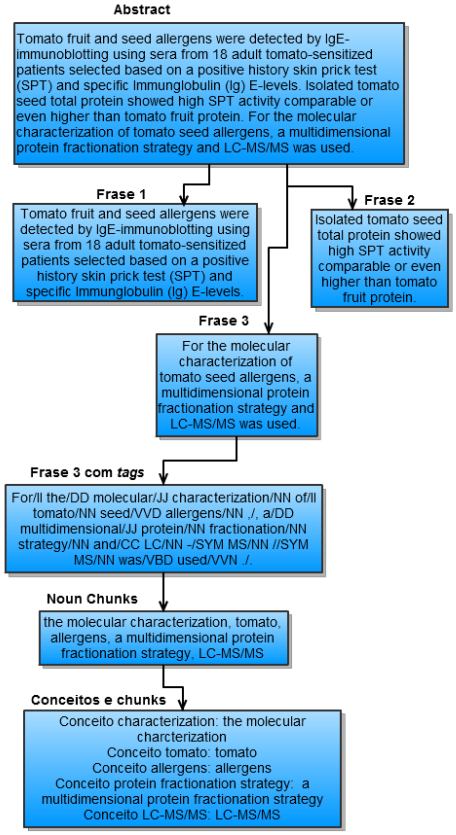

Figura 2- Exemplo a partir de um resumo.

A primeira etapa do algoritmo divide os resumos em frases separadas. Em seguida, a análise léxica ou POST (*Part Of Speech Tagging)* determina as classes gramaticais (*tags*) de cada palavra na frase. Termos compostos, como "*h1n1 virus*", são considerados mais informativos que palavras individuais na hora de descrever um texto (Fagan, 1989). Por essa razão, o próximo passo é obter termos compostos. Ainda, como os substantivos são os melhores descritores, o algoritmo seleciona termos com substantivos (*noun chunks*). Para isto, foram definidos padrões sintáticos baseados em substantivos. A partir da análise sintática da frase, e utilizando o *tag* de cada palavra definido na análise POST, trunca-se a frase em *noun chunks* que combinam com os padrões definidos.

Frequentemente, termos compostos tratam do mesmo assunto, apesar de não possuírem a mesma estrutura gramatical ("*the viruses*", "*virus*" e "*by virus*" tratam de *vírus*). Para considerá-los a mesma entidade introduz-se a idéia de *conceito*. Um conceito é um ou mais substantivos que visa agrupar vários termos similares, neste caso ("*virus*" e "*viruses*"). Percebe-se que esses dois conceitos tratam de um mesmo assunto. Para unificar conceitos similares, utiliza-se o algoritmo de Porter (1980) para a redução de uma palavra até sua raiz e, por exemplo, eliminar plurais. No exemplo, o resultado final inclui todos os termos no conceito "*virus*", sendo esta a palavra-chave.

O processo de obtenção de termos é realizado em todos os resumos dos artigos. O conceito que possui o maior número de artigos distintos tem maior importância, pois ele trata de um assunto que é mais abordado em todos os artigos. Os termos são disponibilizados numa *interface Web*, como mostra a Figura 3. Ao clicar em um termo, os resultados da busca são filtrados e aparecem apenas os artigos que tratam deste assunto.

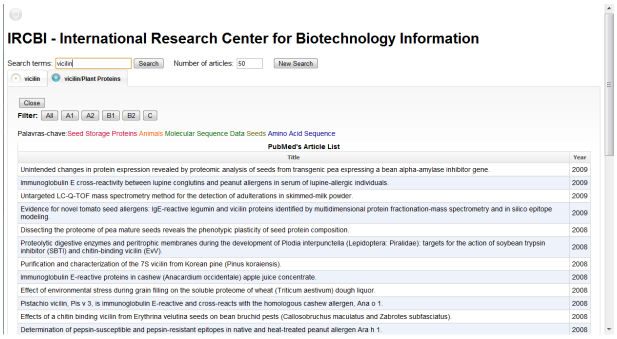

Figura 3. Interface Web

A Figura 4 mostra o fluxo de atividade desde a pesquisa do usuário até a visualização das palavras-chave.

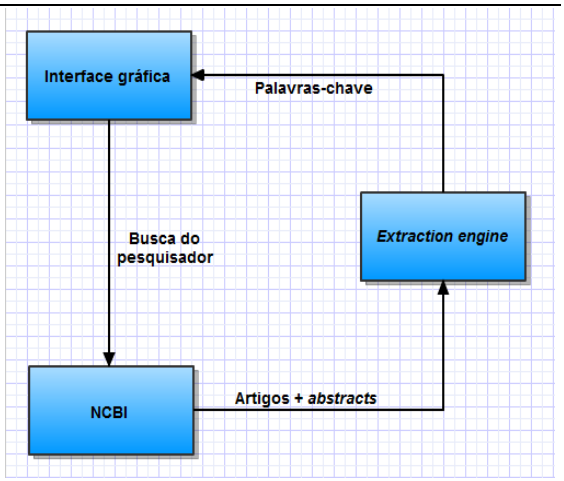

Figura 4- Uma visão geral do processo de obtenção de palavras-chave.

Para a implementação do algoritmo foi utilizada a linguagem Java. A biblioteca de processamento de (LINGPIPE, 2010), que permite a detecção de classes gramaticais de palavras utilizando o MedPOST (SMITH, 2004). Para seu funcionamento, um conjunto de textos manualmente rotulados serve como dados de treinamento (*corpora*). O MEDLINE *corpus* foi elaborado, a partir de artigos do PubMed e utilizado neste trabalho como o *corpora*.

Resultados

Para avaliar a efetividade dos resultados foram utilizadas as duas métricas mais bem aceitas na comunidade de *Information Retrieval*: *precision* e *recall*, além do *fallout*.

A precisão é a proporção de palavras-chave recuperadas que são relevantes para a busca. O *recall* é a proporção de palavras-chave que são relevantes para a consulta e que são recuperadas com sucesso. O f*allout* é a fração de palavras-chave que não são relevantes para a busca e que são recuperadas. (MAKHOUL, 1999).

O intuito inicial foi avaliar a efetividade da análise léxica para determinar as classes

gramaticais das palavras e, principalmente, da obtenção *dos noun chunks*.

O primeiro passo foi escolher dez (10) resumos do PubMed. De cada resumo foram manualmente selecionados os *noun chunks* que deveriam ser detectados pelo *extraction engine*. Estes *noun chunks* são chamados de palavras-chave relevantes. Os resultados obtidos ao executar o *extraction engine*, são chamados de palavras-chave recuperadas. Com esses resultados, foram calculados os valores das métricas de precisão, *recall* e *fallout* para cada resumo.

Tabela 1- Valores das métricas na obtenção de *noun chunks*.

Artigo	Precisão	*Recall*	*Fallout*
1	53%	83%	46%
2	66%	85%	33%
3	67%	94%	32%
4	63%	85%	36%
5	71%	90%	28%
6	77%	88%	22%
7	86%	89%	13%
8	73%	97%	13%
9	71%	93%	28%
10	68%	66%	32%
Média Aritmética	70%	87%	30%

Discussão

Observa-se através dos resultados da Tabela 1 que o algoritmo detecta aproximadamente 87% dos termos desejados, e 70% dos termos recuperados são relevantes. Estes valores são aproximados, pois existem casos em que os *noun chunks* manualmente selecionados podem não ser idênticos aos que o algoritmo detecta. Por exemplo, manualmente foi selecionado o termo literal *"the protein condition"*, mas o *extraction engine* extraiu o termo *"protein condition"*. Isto acarreta uma pequena imprecisão na avaliação dos resultados, pois termos muito parecidos não são detectados.

Para contornar este problema, está sendo usada uma técnica que calcula a similaridade entre termos, chamada *distância de Jaccard*. Esta distância é definida como o tamanho da interseção dividido pelo tamanho da união dos conjuntos de termos representantes de ambos os textos. Neste trabalho, os conjuntos são os *noun chunks*, que possuem uma ou mais palavras.

Conclusão

Os resultados obtidos para a extração de termos foram satisfatórios segundo as métricas de avaliação. Porém, ainda há a necessidade de se testar os resultados obtidos junto a um profissional na área de biotecnologia, que seja capaz de determinar palavras-chave relevantes a um conjunto de artigos. Os próximos passos para a continuação desta pesquisa incluem avaliar a efetividade da extração dos conceitos e o estudo de outras métricas de avaliação de resultados, como a MAP (*Mean Average Precision*), que leva em consideração o *ranking* dos artigos recuperados.

Agradecimentos

CNPq, UENF, LingPipe.

Referências

FAGAN J.L.; The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. **Journal of the American Society for Information Science**, 40(2): 115–139, 1989.

LINGPIPE Home. 2010. Disponível em <http://alias-i.com/lingpipe/>. Acesso em: 19 abr. 2010.

MAKHOUL, J., KUBALA F., SCHWARTZ R., WEISCHEDEL, R Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999.

PORTER M.F., 1980, An algorithm for suffix stripping, **Program**, 14(3): 130−137.

SMITH, L., Rindflesch, T., Wilbur, W.J. MedPost: a part-of-speech tagger for bioMedical text. **Bioinformatics Applications Note** 20(14): 2320-2321, 2004. Disponível em <http://bioinformatics.oxfordjournals.org/cgi/reprint/20/14/2320>. Acesso em: 19 abr. 2010.

WITTEN I.H., PAYNTER G.W., FRANK E., GUTWIN C., NEVILL-MANNING C.G., KEA: Practical Automatic Keyphrase Extraction, 1999.