

Desafios no Processamento de textos de Biomedicina

Rodrigo Rafael Villarreal Goulart¹ e Vera Lúcia Strube de Lima¹

¹PPGCC - Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

rodrigo.goulart@pucrs.br, vera.strube@pucrs.br

Abstract. *The growth of the Biomedical research makes it increasingly difficult for researchers to identify, synthesize and utilize these results in their own researches. This work introduces these problems and how they are being investigated by the Natural Language Processing community.*

Resumo. *O crescimento da pesquisa em Biomedicina dificulta a identificação, síntese e utilização dos resultados pelos pesquisadores. Este trabalho introduz esses problemas e o modo como eles estão investigados pela comunidade de pesquisadores do Processamento da Linguagem Natural.*

Palavras-chave: Processamento de textos de Biomedicina, mineração de textos, entidades nomeadas.

1. Introdução

Estudos como (Ananiadou, 2006; Hunter, 2006; Jin, 2006) colocam que o volume e as características particulares do texto científico da área de Biomedicina dificulta o acesso às informações que atualmente estão disponíveis em artigos acadêmicos via Internet.

Sites de busca tradicionais como o *Google*¹ classificam seus resultados pela popularidade e/ou autoridade que estes representam, mas no domínio das ciências todos os resultados de uma busca podem ser relevantes ou úteis na elaboração de novas hipóteses. Buscar, classificar e compreender as informações disponíveis em artigos científicos, atualmente numerosos, tornou-se uma tarefa exaustiva e complexa para ser executada manualmente, mesmo que esta inclua o uso de *sites* de busca. Este fato vem impulsionando o desenvolvimento de ferramentas para extração automática de informações na área mencionada.

Os experimentos nas áreas de Biologia e Biomedicina fazem uso de dados factuais (extraídos de análises em laboratório) e modelos computacionais que resultam em novas hipóteses sobre seu funcionamento, além das implicações dessas descobertas. Essas informações reúnem dados e conceitos para divulgação utilizando uma terminologia especializada. Atualmente, essas áreas fazem amplo uso dos meios eletrônicos para publicação da sua produção científica.

¹ [Http://www.google.com.br](http://www.google.com.br)

Informações, hipóteses e resultados experimentais apresentados em textos em língua natural são disponibilizados em revistas eletrônicas e *sites* especializados. Grandes bases de dados, com o seqüenciamento de genes ou taxonomias e léxicos, também estão disponíveis na Internet. A base MEDLINE (disponível em <http://www.pubmed.gov/>) e o *site* PubMedCentral (<http://www.pubmedcentral.nih.gov/>) são exemplos relacionados à Biomedicina.

A base MEDLINE é uma base de dados da literatura médica e biomédica compilada e mantida pela Biblioteca Nacional de Medicina dos Estados Unidos. Ela possui cerca de 17 milhões de referências a artigos de cerca de cinco mil veículos de publicação científica. O conteúdo remete a textos desde o ano de 1960 até o presente². Para consultar essa base um sistema de busca chamado PubMed permite a elaboração de pesquisas por tópicos, autores e veículos, além de termos livres (palavras-chave). O resultado de uma pesquisa é uma lista de títulos de artigos, resumos e *links* para os seus respectivos editores.

O *site* PubMedCentral é, além de uma base de referências, um repositório de artigos completos com livre acesso. O objetivo da sua criação foi combater a crescente limitação do acesso a textos completos (os *links* da MEDLINE remetem aos artigos completos geralmente mediante pagamento), resultante de um movimento chamado "Open Access" para o acesso livre a publicações científicas (Hunter, 2006), com cerca de 450 revistas indexadas.

Outro tipo de repositório disponível é a ontologia *Gene Ontology* (GO, <http://www.geneontology.org/>) e a enciclopédia de genes e genomas *Kyoto Encyclopedia of Genes and Genomes* (Kegg, <http://www.genome.jp/kegg/>). Estes repositórios disponibilizam informações na Internet para consulta de sobre termos e dados de sistemas biológicos, respectivamente.

O esforço da GO teve início com a colaboração de três projetos em 1998: o *FlyBase* (<http://flybase.bio.indiana.edu/>), base de dados sobre genes e genomas da mosca *Drosophila*, *Saccharomyces cerevisiae Database* (<http://www.yeastgenome.org/>), base de dados molecular e genética do levedo *Saccharomyces cerevisiae*, e o *Mouse Genome Informatics* (<http://www.informatics.jax.org/>), base de dados sobre genes, genomas e outros dados biológicos sobre ratos. A GO mantém um vocabulário controlado que descreve o produtos dos genes em termos das associações de processos biológicos, componentes celulares e funções moleculares, independentemente de espécies. A enciclopédia Keeg é um conjunto de bases sobre sistemas biológicos que reúne informações sobre redes metabólicas, proteínas, genes, hierarquias e relacionamentos entre objetos biológicos.

A disponibilidade de grandes repositórios de artigos científicos tornou possível a ampla busca de correlações entre proteínas, associações entre doenças e genes e, conseqüentemente, a geração de hipóteses para o desenvolvimento de novos estudos, medicamentos, tratamentos e aplicações em outras áreas. Contudo, analisar manualmente essas informações é uma árdua tarefa que exige do pesquisador a habilidade de identificar e compreender os termos presentes no texto, tarefa que pode ser tornar complexa, como é apresentado a seguir.

² Dados extraídos de http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update em 6 de Junho de 2008

Este trabalho reúne informações sobre a pesquisa na área do Processamento da Linguagem Natural e destaca os desafios na identificação automática de informações contidas em artigos científicos, em especial os da Biomedicina. A próxima seção descreve algumas particularidades do texto de Biomedicina e a seção 3 apresenta estudos sobre o reconhecimento automático de termos desta área. Por fim, a seção 4 expõe uma reflexão a respeito dos desafios apresentados.

2. As particularidades do texto

Segundo Bodenreider (2006), o domínio biomédico tem longa tradição em coletar e organizar termos, como também construir classificações, desde o século XVII.

O emprego de termos especializados pode ser verificado em todas as seções de artigos científicos de Biologia e Biomedicina. O título "*Regulatory networks that function to specify flower meristems require the function of homeobox genes PENNYWISE and POUND-FOOLISH in Arabidopsis*" (Kanrar, 2008) exemplifica o emprego de terminologia especializada na determinação de espécie (*Arabidopsis*), genes (*PENNYWISE and POUND-FOOLISH*) e processos biológicos (*meristem*). O leitor com a capacidade de identificar e compreender o significado dos termos tem a oportunidade de reunir documentos por assunto e estabelecer relacionamentos entre os conceitos contidos nos mesmos. No entanto, a grande quantidade de textos e neologismos dificulta a análise manual, seja pelo surgimento novos termos ou de novos significados levando a ambigüidade, que apresenta desafios específicos na compreensão de termos biomédicos. Os múltiplos significados entre genes polissêmicos, de acordo com experimentos realizados por Chen (2005), chegam a 14,2% entre espécies. A polissemia pode ser ilustrada pelo termo *NF2*. Apenas as letras em maiúsculo e minúsculo servem para diferenciar o gene humano (*NF2*) daquele de ratos (*Nf2*). Além disso, a expressão *NF2* é simultaneamente o nome de um gene, a proteína que ele produz, e a doença resultante da sua mutação.

Meios para desambiguação estão sendo investigados (Kim, 2003; Liu, 2004; Roberts, 2008; Wang, 2008) mas o novas estratégias para a especificidade dos termos em Biologia Molecular (i.e. ambigüidade entre espécies) são um tema em aberto (Bodenreider, 2006).

O estudo da variabilidade ortográfica, léxica e semântica de termos é uma das questões em aberto que motiva a pesquisa na área do Processamento da Linguagem Natural (PLN), por influenciar diretamente o desempenho de procedimentos automatizados de busca em textos de Biomedicina.

3. Reconhecimento de termos

O reconhecimento de termos é tema de pesquisa antigo na área de PLN. Dicionários, sistemas de regras e aprendizado de máquina são exemplos de técnicas investigadas em domínios específicos do conhecimento.

Para Baumgartner (2008) e Park (2006) as abordagens para o reconhecimento de entidades biológicas, em especial genes e proteínas, podem fazer uso de dicionários, regras, aprendizado de máquina e abordagens híbridas. O aprendizado de máquina, de acordo com Yang (2008), é o mais popular por identificar termos ausentes nos

dicionários. Por outro lado, o aprendizado de máquina carece de informações adicionais sobre os termos reconhecidos como, por exemplo, as relações homólogas entre termos relacionados a mamíferos.

Abordagens baseadas em dicionários podem ser úteis se o termo procurado for encontrado no dicionário, mas o seu desempenho está associado ao tamanho e à qualidade do repositório. Dicionários podem ser constituídos de bases disponíveis na Internet, como os dicionários *Mouse Genome Database MGD* (<http://www.informatics.jax.org/>), *UniProt* (<http://www.pir.uniprot.org/>) e *National Center for Biotechnology Information NCBI Entrez Taxonomy* (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy>). A vantagem em utilizar dicionários *on-line* é que novos nomes e símbolos são adicionados pelos grupos de pesquisa que os mantêm.

Por outro lado, um dos problemas enfrentados na utilização de dicionários é a variação ortográfica dos termos, que pode gerar falsos positivos, resultando na associação de termos ambíguos, ou falsos negativos, quando não são reconhecidos sinônimos, variações e novos nomes de uma mesma entidade. Por exemplo, a proteína *NF-Kappa B* pode ser encontrada nas grafias *NF Kappa B*, *NF KappaB* ou *NFKappaB*.

Ananiadou (2006), relaciona os seguintes desafios na manutenção de repositórios terminológicos em Biomedicina :

1. Reconhecimento preciso dos limites e da composição ortográfica de termos para identificação e classificação do mesmos;
2. Tratamento eficiente das variações de termos;
3. Resolução de anáforas e outros relacionamentos entre termos co-referentes, para o estabelecimento de relações entre termos;
4. A seleção dos termos e conceitos mais significativos de um documento para a indexação e recuperação de informações;
5. Estudos sobre o reconhecimento de entidades para outras classes (hoje concentradas basicamente em proteínas e genes) para auxiliar procedimentos de mineração de textos e dados.

No aprendizado de máquina a identificação de um termo pode ser realizada por métodos supervisionados e não supervisionados.

Collier (2000) utiliza como método o aprendizado supervisionado. Como conjunto de treino foi utilizado o corpus da base MEDLINE (100 resumos) com trechos das sentenças cujos nomes de proteínas DNA foram anotados. O modelo de aprendizado utilizado é *Hidden Markov Model* (HMM), treinado por meio de bigramas (pares de palavras). Os atributos para o aprendizado consistem na identificação da presença ou ausência de caracteres especiais (letras em maiúsculo ou minúsculo, dígitos, etc) ou no fato da palavra em análise ser um determinante (ex. *The*) ou uma conjunção (ex. *and*). O modelo determina para cada termo anotado a probabilidade de uma palavra pertencer a ele. Por fim, o conjunto de teste do aprendizado determina a seqüência de anotações mais provável para uma seqüência de palavras de entrada. De acordo com Bodenreider (2006) ainda há pouco corpora anotado, a exemplo do corpus GENIA ([GT – Lingüística e Computação](http://www-</p></div><div data-bbox=)

tsujii.is.s.u-tokyo.ac.jp/GENIA/), o que limita o desenvolvimento de métodos supervisionados.

Por sua vez, Morgan (2004) apresenta um método para construir corpora anotado em grande quantidade utilizando uma base de dados com nomes de genes mantida manualmente, a FlyBase³. Os nomes são extraídos, por meio de expressões regulares, de resumos da base MEDLINE que também estão referenciados na base. Desta forma, os nomes de genes (e seus sinônimos) são anotados nos resumos por um processo automático (com 78% de precisão, 88% de abrangência e um F-measure de 83%). Esse processo de anotação de novos textos é utilizado no aprendizado de um HMM para o reconhecimento de nomes de genes, cuja precisão alcança 78%, uma abrangência de 71% e um F-measure de 75%.

Soluções híbridas incluem a utilização simultânea de dicionários, regras ou aprendizado de máquina, com a finalidade de aumentar a precisão e abrangência da identificação de termos.

Em Yang (2008) propõe-se um método para minimizar os problemas com abreviações de termos, que inclui três etapas: a construção e expansão de um dicionário de entidades biológicas (termo e abreviação), um método de aproximação de *strings*, e o pós-processamento para identificação e associação de novas abreviações com termos de um dicionário. Os experimentos realizados por Yang na construção e expansão do dicionário, reconhecimento de entidades por similaridade e pós-processamento, utilizaram o corpus do JNLPBA (*Joint Workshop on Natural Language Processing in Biomedicine and its Applications*), associado ao COLING 2004⁴, com as entidades e suas respectivas classes (*protein, RNA, DNA, cell line, cell type*) etiquetadas. Os melhores resultados obtiveram um F-score de 68,8%. Os erros mais comuns foram os termos não reconhecidos (54% do total de erros).

4. Considerações

A perspectiva de desenvolver sistemas que auxiliem os pesquisadores das áreas biológicas e biomédicas na elaboração de novas hipóteses é uma das finalidades mais avançadas e promissoras que se apresentam, quando se vislumbra o uso de técnicas de PLN nessa área.

Dada a sobrecarga de dados e informações, os trabalhos estudados remetem ao fato de que não será possível desenvolver inovações tecnológicas nessas áreas, num curto espaço de tempo, sem o emprego de mecanismos automáticos que minerem as informações disponíveis. Além da quantidade de artigos disponíveis, o número e a variabilidade dos termos dificulta a compreensão dos textos, tanto pelos leitores como pelos mecanismos de busca. A variabilidade ortográfica, léxica e semântica são questões que conduzem o interesse na compreensão automática de textos por influenciar diretamente os resultados da extração de informações. Sem o reconhecimento preciso dos termos e de seus significados não é possível classificar documentos ou estabelecer relacionamentos entre os conceitos contidos nos mesmos.

³ <http://flybase.org/>

⁴ <http://www.issco.unige.ch/coling2004/>

O reconhecimento de termos é tema de pesquisa antigo na área de PLN. O emprego de dicionários, sistemas de regras e aprendizado de máquina são exemplos de técnicas empregadas em domínios do conhecimento específicos. Dicionários são ricos em informações sobre os termos relacionados mas são limitados quanto à generalização. Sistemas de regras e aprendizado de máquina conseguem se adaptar melhor às variações léxicas e sintáticas mas são dependentes da sua manutenção ou da quantidade de informações disponível para sua construção.

Dentre os eventos destacados por autores da área (Baumgartner, 2008; Hirschman, 2002; Hunter, 2006; Park, 2006 e Yang, 2008), o workshop BioNLP (associado à conferência ACL) e a track Genomics da conferência em recuperação de informação TREC (<http://trec.nist.gov/>) são os mais citados e levantam a necessidade de desenvolver métodos para extração de informações (reconhecimento e extração de relacionamentos entre entidades), construção de repositórios ou bases de conhecimento para o domínio de Biologia e Biomedicina, além de métodos para avaliação dos benefícios, resultados e necessidades para a pesquisa nessa área.

5. Referencias e Citações

- ANANIADOU, S.; NENADIC, G. Automatic Terminology Management in Biomedicine Text Mining for Biology and Biomedicine, Artech House Books, 2006, 67-98
- BODENREIDER, O. Lexical, Terminological, and Ontological Resources for Biological Text Mining Text Mining for Biology and Biomedicine, Artech House, 2006, 43-66
- CHEN, L.; LIU, H.; FRIEDMAN, C. Gene name ambiguity of eukaryotic nomenclatures Bioinformatics, Oxford Univ Press, 2005, 21, 248-256
- COLLIER, N.; NOBATA, C.; TSUJII, J. Extracting the names of genes and gene products with a hidden Markov model Proceedings of the 18th conference on Computational linguistics, Association for Computational Linguistics, 2000, 201-20
- HIRSCHMAN, L.; MORGAN, A.; YEH, A. Rutabaga by any other name: extracting biological names Journal of Biomedical Informatics, Elsevier, 2002, 35, 247-259
- HUNTER, L.; COHEN, K. Biomedical Language Processing: What's Beyond PubMed? Molecular Cell, Elsevier, 2006, 21, 589-594
- JIN, Y. et al. Automated recognition of malignancy mentions in biomedical literature BMC Bioinformatics, BioMed Central, 2006, 7, 492
- KANRAR, S. et al. Regulatory networks that function to specify flower meristems require the function of homeobox genes PENNYWISE and POUND-FOOLISH in Arabidopsis The Plant Journal, 2008, 54, 924-937

- KIM, J. et al. GENIA corpus-a semantically annotated corpus for bio-textmining Bioinformatics, Oxford Univ Press, 2003, 19, 180-182
- LIU, H.; TELLER, V.; FRIEDMAN, C. A multi-aspect comparison study of supervised word sense disambiguation Journal of the American Medical Informatics Association, Elsevier, 2004, 11, 320-331
- MORGAN, A. A. et al. Gene name identification and normalization using a model organism database Journal of Biomedical Informatics, Named Entity Recognition in Biomedicine, 2004, 37, 396-410
- PARK, J. C.; KIM, J. Named Entity Recognition Text Mining for Biology and Biomedicine, Artech House, 2006, 43-66
- ROBERTS, Angus et al. Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008
- WANG, X.; GROVER, C. (ELRA), E. L. R. A. (ed.) Learning the Species of Biomedical Named Entities from Annotated Corpora Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008
- YANG, Z.; LIN, H.; LI, Y. Exploiting the contextual cues for bio-entity name recognition in biomedical literature Journal of Biomedical Informatics, Elsevier, 2008