O Contexto no Reconhecimento de Entidades Nomeadas em Textos de Biomedicina

Rodrigo R. V. Goulart[1] e Vera L. Strube de Lima[2]

[1]Instituto de Ciências Exatas e Tecnológicas – Centro Universitário Feevale
RS-239, 2755 - Novo Hamburgo - RS – Brasil.

[2]Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681 - Prédio 32 - Sala 647 - Porto Alegre - RS – Brasil

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

Abstract. *This article presents a study on Named Entities (NE) recognition using contextual information present on a Biomedical corpus. Related work indicates that the use of context (words surrounding a word) can assist the NE recognition. This work presents experimental results to evaluate the impact of different context settings, using machine learning, for the NE recognition.*

Resumo. *Este artigo apresenta um estudo sobre a utilização de informações contextuais na identificação de Entidades Nomeadas (EN) presentes num corpus de Biomedicina. Trabalhos relacionados indicam que o emprego do contexto (palavras no entorno de outra palavra) pode auxiliar o processo de reconhecimento de ENs. Este trabalho apresenta os resultados de experimentos com a finalidade de avaliar o impacto de diferentes configurações de contextos, no aprendizado de máquina, para o reconhecimento de ENs.*

1. Introdução

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

O interesse em processar textos dessa área se dá, por exemplo, pela necessidade de viabilizar mecanismos para recuperação de documentos e extração de informações contidas nesses textos. Hipóteses, experimentos e resultados descritos em artigos científicos exigem do leitor conhecimento especializado para identificar, compreender e relacionar as informações contidas nos mesmos. Neste cenário, o Reconhecimento de Entidades Nomeadas (REN) em domínios especializados é um fator crucial para a compreensão dos textos.

Entidades Nomeadas (EN) são termos especializados como, por exemplo no domínio da Biomedicina, o nome "NF2" (que se refere a um gene). No entanto identificar e compreender o significado de um termo não é uma tarefa trivial. No exemplo anterior apenas as letras em maiúsculo e minúsculo servem para diferenciar o gene humano (NF2) daquele de ratos (Nf2). Além disso, segundo Chen et al. (2005), a expressão "NF2" é simultaneamente o nome de um gene, a proteína que ele produz, e a doença resultante da sua mutação. Os múltiplos significados de ENs de genes polissêmicos, de acordo com experimentos realizados por Chen, chegam a 14,2% entre

espécies.

A quantidade e a variedade de termos e informações relacionados a genes e proteínas é tão expressiva que bases terminológicas têm sido disponibilizadas e mantidas para organizar esses termos. Essas bases podem ser empregadas na recuperação de documentos e informações por meio dos relacionamentos semânticos entre termos que elas mantêm. A Gene Ontology (GO, http://www.geneontology.org/), por exemplo, é uma base de dados que mantém cerca de 27 mil termos sobre genes e proteínas[1]. Nela podem ser encontradas definições, relacionamentos entre termos ou relacionamentos com outras bases de dados. Raychaudhuri (2006) registra que, somente entre junho de 2003 e junho de 2004, 3.652 termos foram adicionados à GO. Por outro lado, o volume de artigos científicos de Biomedicina também está crescendo, o que torna as bases terminológicas rapidamente desatualizadas. A base de artigos de Biomedicina PUBMED recebeu entre 2005 e 2008 cerca de 4 milhões de novos artigos[2], o que demonstra a necessidade de obter meios para extrair novas ENs presentes nesses artigos.

Dentre as alternativas para identificação automática de ENs, a utilização do contexto (palavras no entorno de uma palavra ou conjunto de palavras) é um dos métodos para diferenciar termos que são ENs de um determinado domínio das que não fazem parte do mesmo. Este trabalho apresenta experimentos e resultados sobre o impacto de diferentes configurações de contexto, no aprendizado de máquina, para o reconhecimento de ENs. Para isso, são levantadas duas hipóteses: (1) o contexto produz melhores resultados para determinadas categorias de ENs (ex. proteínas, genes, etc); (2) a quantidade de palavras consideradas no entorno influencia o REN.

A Seção 2 relaciona diferentes métodos para o reconhecimento de ENs e trabalhos correlatos. A Seção 3 descreve a metodologia empregada: o corpus utilizado, a preparação, modelagem e execução dos experimentos. A Seção 4 apresenta os resultados e, finalmente, a Seção 5 apresenta considerações sobre esses resultados.

2. O reconhecimento de ENs em textos de Biomedicina

Gene Ontology, OBO Foundry (http://www.obofoundry.org/), MeSH (http://www.nlm.nih.gov/mesh/) e UniProt (http://www.uniprot.org/) são exemplos de bases terminológicas sobre genes e proteínas. Reconhecer os nomes das classes de termos nelas contidas é uma etapa importante na tarefa de recuperação de informações. No entanto coletá-los e organizá-los não é uma tarefa simples. Por exemplo, considere as sentenças: (1) "*The gene _____ is expressed under …"* (2) "*... which induces _____ expression, …"*. Por meio da presença e localização dos termos *gene* e *expressed*, na sentença (1), o leitor é capaz de concluir que o termo omitido no espaço em branco é um gene. A sentença (2) exige do leitor um vocabulário e conhecimento mais especializado, fato decorrente da presença dos termos *induces* e *expression.* Estes formam o contexto do termo procurado.

Abordagens computacionais, a exemplo dos especialistas humanos, também podem fazer uso do contexto no reconhecimento de ENs. Neste caso, o contexto pode ser definido pelos termos à esquerda e à direita do termo, compreendendo uma "janela"

1 Quantidade extraída de http://www.geneontology.org/GO. downloads.ontology.shtml em 05/03/2009.

2 Em (Raychaudhuri, 2006) estima-se em 14 milhões de artigos no ano de 2005 e, de acordo com o site http://www.ncbi.nlm.nih.gov/pubmed/, são cerca 18 milhões em 27/04/2009. Não foram encontrados dados históricos na base PUBMED.

de termos antes e após ou, até mesmo, todos os termos de uma dada sentença. A estratégia mais simples é considerar os termos imediatamente à esquerda e à direita, e verificar se estes termos combinam com alguma palavra-chave pré-estabelecida. Os modelos probabilísticos ou baseados em regras, por exemplo, são meios de diferenciar o contexto de termos que fazem referência a ENs dos que não o fazem.

Para Raychaudhuri (2006) e Hahn (2006) existem várias abordagens metodológicas que podem ser empregadas para estabelecer o conjunto de *features*[3] no reconhecimento de nomes de genes e proteínas, dentre as quais podemos destacar: reconhecer nomes por meio de dicionários; utilizar as características ortográficas; utilizar características sintáticas; avaliar o contexto de um termo; utilizar características morfológicas; tratar nomes de genes e suas abreviaturas.

Muitos trabalhos fazem uso de abordagens híbridas (Laws, 2008; Kim et al., 2004; GuoDong et al., 2004; Finkel et al., 2004 e Settles et al., 2004), ou seja, empregam mais de uma metodologia simultaneamente na tentativa de obter melhores resultados. No Joint Workshop of BioNLP/NLPBA (Kim et al., 2004) os participantes propuseram diferentes *features* e algoritmos para o aprendizado de máquina na identificação de ENs de diferentes categorias (proteínas, DNA, RNA, etc). A Tabela 1 relaciona os F-*scores* dos três primeiros colocados, os resultados parciais de cada categoria, os resultados finais, assim como as *features* e algoritmos empregadas pelos mesmos. Esses resultados serão úteis na compreensão deste trabalho.

Tabela 1: F-*score* e técnicas empregadas no JNLPBA de 2004 (Kim et al., 2004)

	GuoDong e Jian (2004)	**Finkel et al. (2004)**	**Settles (2004)**
Protein	**73.77**	72.67	72.00
RNA	64.10	**68.83**	64.70
Overall	**72.55**	70.06	69.50
features	Ortografia, sintaxe, dicionários e morfologia	Ortografia, sintaxe, dicionários e contexto	Ortografia e dicionários
Algoritmos de aprendizado[4]	HMM e SVM	MEMM	CRF

Kim (2004) destaca que a quantidade de *features* teoricamente influencia os resultados. Esse fato pode ser observado no trabalho de GuoDong (2004), que obteve os melhores resultados. No entanto, a performance de GuoDong no reconhecimento de termos da categoria RNA não foi a melhor (F-*score*=64.10). Por outro lado, o resultado (F-*score*=68.83) obtido por Finkel (2004) pode estar relacionado com o fato de ter sido empregado o contexto dos termos na classificação, *feature* não utilizada por GuoDong. Uma possibilidade é que haja algum relacionamento entre ENs da categoria de RNAs e os termos do seu entorno, que auxilie a sua identificação.

3. O experimento com o contexto de proteínas e RNAs

Ao invés de utilizar diferentes tipos de *features* (ortografia, sintaxe, etc), este trabalho propõe a utilização de apenas um tipo de *feature*, a de contexto, onde cada palavra do

3 Características dos termos, contextos ou documentos, por exemplo, que são empregados no aprendizado de máquina.

4 HMM – Hidden Markov Model, SVM – Support Vector Machine, MEMM – Maximum Entropy Markov Model e CRF – Conditional Random Fields.

contexto representa uma *feature*. O objetivo é investigar as hipóteses levantadas na Seção 1 utilizando o corpus GENIA (apresentado na Seção 3.1.), que foi elaborado para o Workshop BioNLP/NLPBA, utilizando a categoria mais representativa *(protein)* e a menos representativa (RNA). A seguir são apresentados detalhes do corpus GENIA, a modelagem do experimento e os resultados obtidos.

3.1. Corpus GENIA

Neste trabalho foi utilizado o corpus GENIA[5], que é uma coleção de 2.404 artigos (título e resumo) da base de artigos de Biomedicina Medline[6] produzido para o Workshop BioNLP/JNLPBA de 2004.

O corpus de treino é composto por 472.006 palavras e 20.546 sentenças. Nele estão anotadas 51.301 ENs (109.588 palavras) divididas em 36 sub-categorias das classes de termos *protein*, DNA, RNA, *cell line* e *cell type*. A categoria *protein* é representada por 30.269 ENs (55.117 palavras) e RNA com 951 ENs (2.481 palavras). As categorias *protein* e RNA são a mais e a menos representativas, respectivamente. O corpus de teste é composto por 96.780 palavras. Nele estão anotadas 8.662 ENs (19.392 palavras) utilizando as mesmas categorias do corpus de treino. A categoria *protein* é representada por 5.067 ENs (9.841 palavras) e RNA com 118 ENs (305 palavras).

Um exemplo de anotação contido no texto é o da sentença apresentada no Quadro 1. Nessa sentença cada palavra é anotada com um código que determina se ela não faz parte de uma EN (O – *other*), se ela é a primeira palavra de uma EN (B – *begin*) ou outra palavra da EN que não seja a inicial (I - *in*). Algumas ENs podem ser simples (ex.: *IFN-gamma)* ou compostas (ex.: *TNF mRNA*). Além da anotação que identifica ENs, o corpus possui outros níveis de anotação não presentes nessa versão (Kim et al., 2004).

3.2. A modelagem do experimento

O experimento compreende duas etapas: treino e teste. A etapa de treino utiliza o corpus de treino do GENIA e executa os procedimentos apresentados na Figura 1 e descritos no texto que segue.

O primeiro procedimento é extrair os artigos do corpus original (procedimento 1, Figura 1) por meio de identificadores contidos no arquivo original. Em seguida a frequência TF-iDF (Lavelli, 2004) das palavras (2) é calculada, com a finalidade de determinar a importância de cada termo de um documento em relação ao corpus. Com as frequências das palavras o procedimento é de extração dos exemplos positivos e negativos de ENs (3) para dada classe (i.e. *protein* ou *RNA)* de acordo com um tipo de janela. Neste experimento foram analisadas janelas de 1, 2, 3 ou 4 palavras do contexto (ex. 2 palavras à esquerda e à direita).

Priming	*O*	*manifested*	*O*	*TNF*	*B-RNA*
by	*O*	*at*	*O*	*mRNA*	*I-RNA*
IFN-gamma	*B-protein*	*the*	*O*	*accumulation*	*O*
was	*O*	*level*	*O*	*.*	*O*
primarily	*O*	*of*	*O*		

Quadro 1: Exemplo de sentença do corpus GENIA

5 Disponível em http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/
6 Disponível em http://www.ncbi.nlm.nih.gov/pubmed/

Além disso, as ENs são tratadas do ponto de vista das palavras e não dos termos. Isso significa que são consideradas as palavras, e suas respectivas classes, de cada EN identificada no corpus, seja ela um termo simples (i.e. compreendendo uma única palavra como, por exemplo, *"CD28")* ou composto (contendo mais de uma palavra, como é o caso de "*10-amino acid CD40 cytoplasmic signaling determinant*"). Deste modo, uma palavra pode estar presente duas vezes na lista (como uma das palavras de um termo composto). Por outro lado, ela também pode ser encontrada no texto classificada como *other*. A diferença entre utilizar palavras ou termos é que termos compostos excluem a possibilidade de avaliar a influência das palavras contidas neles para identificação de uma EN. Por exemplo, a palavra *transcripts* contida na EN "*IL-2R alpha transcripts*" também pode ser encontrada no corpus classificada como *other*, mas esse fato não exclui a possibilidade dela fazer parte de outras ENs, como pode ser constatado na EN "*C transcripts*". Essa abordagem tem como objetivo permitir que novas ENs que utilizam *transcripts* possam ser detectadas.

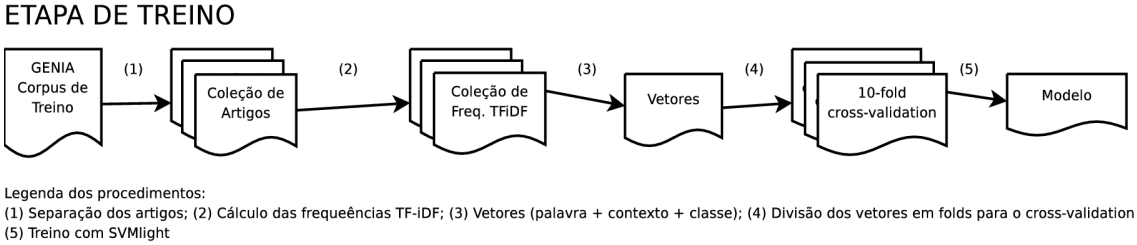

Figura 1: Procedimentos com o corpus de treino do GENIA

Um exemplo positivo ou negativo é chamado de vetor e reúne uma determinada palavra, sua classificação e contexto. Os vetores do Quadro 2 exemplificam uma entrada positiva (+1, para a classe *protein)* e outra negativa (-1), utilizando uma janela de 1 palavra, e as palavras (ex.: "4:" corresponde a palavra *in*) com seu respectivos TF-iDF (ex.: o TF-iDF de "4:" é .000709097) este exemplo corresponde à sintaxe do algoritmo SMVlight (descrito na Seção 3.3.).

+1	4:.000709097	114:.102460172	280:.138992020
-1	4:.000709097	67:.1106112242	114:.102460172

Quadro 2: Exemplos de vetores (positivo e negativo, janela de 1 palavra)

Um arquivo de vetores é então dividido em *folds* (4, Figura 1) com a finalidade de viabilizar a validação do aprendizado utilizando o método de *10-fold cross-validation* (Alpaydin, 2004), que divide o conjunto de vetores em 10 partes para validação cruzada. Esta etapa (5, Figura 1) viabiliza a avaliação do desempenho dos modelos adquiridos com o treino. Em seguida é dado início à etapa de teste, descrita na Figura 2.

O teste utiliza o corpus de teste do GENIA e executa, em grande parte, os mesmos procedimentos do treino (procedimentos 1, 2 e 3 presentes nas Figuras 1 e 2). O diferencial é que um vetor contendo apenas as frequências TF-iDF e apenas um modelo da fase de treino são fornecidos ao classificador do SVMlight. Este produz uma lista (6) das palavras do corpus de treino com as suas respectivas classes (i.e. positivo ou negativo).

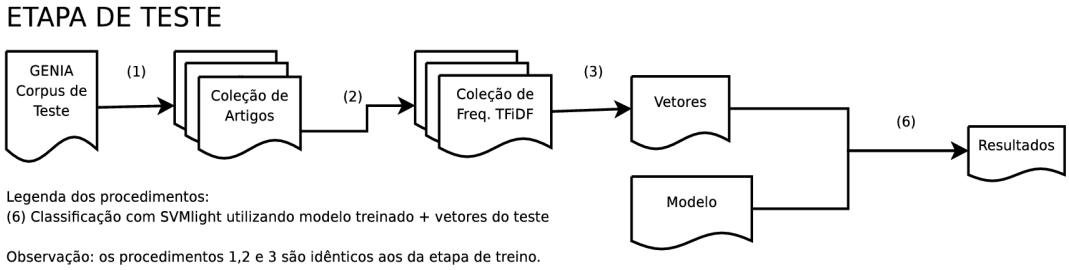

Figura 2: Procedimentos com o corpus de teste do GENIA

3.3. O pré-processamento e o classificador SVM

A implementação dos experimentos envolve o uso de ferramentas para extração de informações e execução de cálculos para produção dos vetores de entrada. Foram utilizados o processador GAWK e scripts shell BASH para Linux. Os resultados do Workshop BioNLP/ JNLPBA em 2004 (Kim, 2004), além de outros trabalhos em RENs (Bin, 2008; Roberts, 2008; Vlachos, 2007) apontam o emprego de *Support Vector Machines* (SVM) como um algoritmo que apresenta uma das melhores performances em REN. A implementação utilizada nestes experimentos é a disponibilizada por Joachims (1999) chamada SVMlight[7]. Foram utilizado o *kernel* linear e parâmetros *default* na execução do programa.

4. RESULTADOS

Os dados fornecidos pelo programa classificador SVMlight relacionam precisão e abrangência. Os resultados compreendem os testes com janelas variando entre 1 a 4 palavras e a classificação de RNAs e proteínas (classe *protein*).

A precisão (P) é determinada pela proporção de casos classificados como positivos que são realmente positivos. A abrangência (Ab) é calculada com base na proporção de casos positivos que foram identificados. A acurácia (Ac) é determinada pela proporção total de classificações corretas (positivas e negativas). Para calcular o F-*score* dos resultados obtidos é empregada a média harmônica utilizada no JNLPBA definida como F = (2.P.Ab) / (P+Ab). A Tabela 2 apresenta os F-*scores* dos testes com os modelos produzidos por cada *fold* da etapa de teste. A média aritmética (M_a), com as variações do desvio padrão (σ) e, finalmente, o coeficiente de variância (C_v). O gráfico 1 expressam as variações dos valores da Tabela 2 em relação às janelas.

Tabela 2: Resultados do teste para (a) RNAs e (b) protein

F-*score*	RNA				
Janela	Ma - σ	Min	Max	Ma - σ	Cv
1	7,01	6,33	**28,25**	23,74	**54,4**
2	0,02	0,49	13,66	9,3	99,55
3	-0,23	**0,29**	8,9	5,99	107,92
4	-0,64	0,37	9,94	6,53	121,71

(a)

F-*score*	*protein*				
Janela	Ma - σ	Min	Max	Ma - σ	Cv
1	5,92	4	**9,37**	9,5	23,24
2	4,46	3,98	8,13	6,84	21,11
3	4,23	4,07	6,63	5,79	**15,58**
4	3,75	**3,25**	7,21	6,27	25,18

(b)

Dadas as hipóteses levantadas e os procedimentos empregados nos experimentos executados neste trabalho, é possível constatar que a categoria que obtém os melhores

7 Disponível em http://svmlight.joachims.org/

resultados é a de RNAs (F-*score*=28,25). A melhor janela de contexto, com 1 palavra, produziu os melhores resultados, tanto para a classe RNA como para *protein*. Os Gráficos 1.a e 1.b mostram que, mesmo com coeficientes de variação altos, os resultados obtidos com os modelos com janelas de 1 palavra são superiores aos obtidos com janelas maiores.

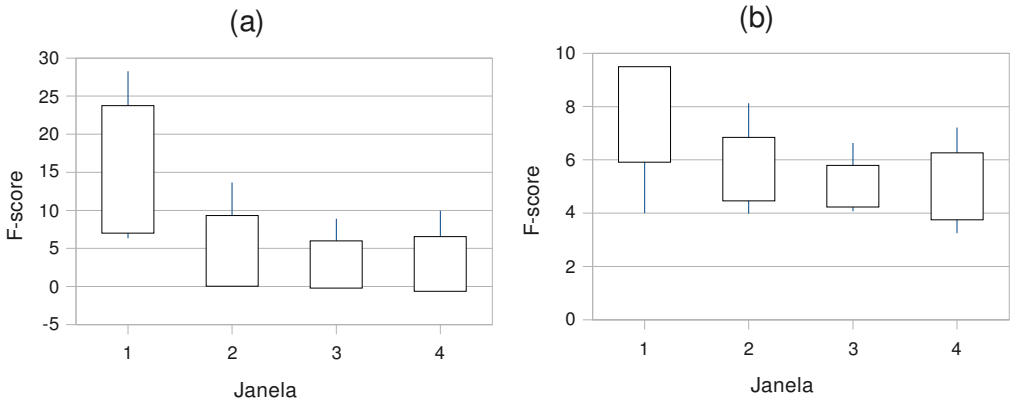

Gráfico 1: Desempenho do teste para (a) RNA e (b) protein

5. CONSIDERAÇÕES

Analisando as 10 ENs mais comuns (Tabela 3), presentes no corpus de treino ou no corpus de teste, predominam as ocorrências de bi-gramas na classe RNA e uni-gramas na classe *protein*. De fato, esta característica é predominante em todo o corpus, de acordo com as ocorrências absolutas dispostas no Tabela 4. Além disso, ressalta-se a ocorrência da palavra "mRNA" como um indicador de um RNA. Considerando os resultados dos experimentos e os n-gramas presentes no corpus é possível inferir que: (i) a identificação de proteínas não obtém bons resultados por meio do contexto, (ii) por outro lado, a identificação RNAs sofre influência do contexto, da influência mútua das palavras de termos compostos e de determinadas formas, como por exemplo "mRNA".

Tabela 3: ENs mais comuns do corpus de treino e de teste para as classes protein e RNA

Corpus de treino				Corpus de teste			
protein		RNA		protein		RNA	
Contagem	EN	Contagem	EN	Contagem	EN	Contagem	EN
861	NF-kappa B	78	mRNA	92	NF-kappaB	5	TNF-alpha mRNA
540	NF-kappaB	21	GR mRNA	70	NF-kappa B	4	MCP-1 mRNA
534	IL-2	21	c-jun mRNA	64	glucocorticoid receptor	4	cytokine mRNA
331	transcription factors	19	mRNAs	62	IL-2	3	RXRalpha mRNA
317	AP-1	18	IL-2 mRNA	53	glucocorticoid receptors	3	IL-6 mRNA
314	IL-4	17	c-fos mRNA	51	IL-4	3	ER mRNA
283	transcription factor	13	IL-6 mRNA	46	GR	3	Egr-1 mRNA
243	TNF-alpha	13	c-myc mRNA	40	cytokines	2	nm23-H1 and -H2 mRNA
226	IFN-gamma	11	C/EBP epsilon mRNA	39	TNF-alpha	2	mRNA
200	cytokines	11	1 , 25 (OH) 2D3 receptor RNA	39	AP-1	2	MIP-2 mRNA

Não nos parece adequado afirmar que estes fatores são determinantes para identificação de ENs, a ponto de empregá-los na forma de regras. Outros trabalhos não empregam apenas regras como uma forma de REN por, conforme dito na Seção 2, não permitirem a adaptação a novos casos com vista à generalização do REN. O emprego de aprendizado de máquina produz modelos de aprendizado que não são passíveis de

prover um entendimento racional das predições para o REN. Por esta razão a compreensão do comportamento deste método no REN, utilizando apenas a *feature* de contexto, é de caráter exploratório e experimental.

Tabela 4: Contagem de ENs que representam termos compostos e simples

	Corpus de treino		Corpus de teste	
	Simples	Compostos	Simples	Compostos
RNA	128	823	16	102
protein	17190	13092	2707	2359

Como trabalhos futuros se ambiciona a utilização de outros tipos de *features*, como a presença de prefixos, sufixos e pontuação, além do emprego de dicionários e informações de categorias gramaticais (POS). Por fim, é pretendido utilizar outros tipos de anotações presentes no corpus GENIA e empregar os resultados deste trabalho na identificação de eventos biológicos, como por exemplo regulação.

Referências

Ananiadou, S.; Kell, D. & Tsujii, J. Text mining and its potential applications in systems biology Trends in Biotechnology, Elsevier, 2006, 24, 571-579.

Bin, C.; Xiaofeng, Y.; Jian, S. & Lim, T. C. Other-Anaphora Resolution in Biomedical Texts with Automatically Mined Patterns Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 2008, 121-128.

Chen, L.; Liu, H. & Friedman, C. Gene name ambiguity of eukaryotic nomenclatures Bioinformatics, Oxford Univ Press, 2005, 21, 248-256.

Finkel, J.; Dingare, S.; Nguyen, H.; Nissim, M.; Manning, C. & Sinclair, G. Collier, N.; Ruch, P. & Nazarenko, A. (ed.) Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web COLING 2004 (NLPBA/BioNLP) 2004, COLING, 2004, 91-94.

GuoDong, Z. & Jian, S. Collier, N.; Ruch, P. & Nazarenko, A. (ed.) Exploring Deep Knowledge Resources in Biomedical Name Recognition COLING 2004 (NLPBA/BioNLP) 2004, COLING, 2004, 99-102.

Hahn, U. & Wermter, J. Levels of natural language processing for text mining Text Mining for Biology and Biomedicine, 2006, 13-41.

Hunter, L. & Cohen, K. Biomedical Language Processing: What's Beyond PubMed? Molecular Cell, Elsevier, 2006, 21, 589-594.

Jin, Y.; McDonald, R.; Lerman, K.; Mandel, M.; Carroll, S.; Liberman, M.; Pereira, F.; Winters, R. & White, P. Automated recognition of malignancy mentions in biomedical literature BMC Bioinformatics, BioMed Central, 2006, 7, 492.

Joachims, T. Making large-scale support vector machine learning practical MIT Press Cambridge, MA, USA, 1999.

Kim, J.; Ohta, T.; Tsuruoka, Y.; Tateisi, Y. & Collier, N. Introduction to the bio-entity recognition task at JNLPBA Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04), 2004, 70-75.

Lavelli, A.; Sebastiani, F. & Zanoli, R. Distributional term representations: an experimental comparison Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004, 615-624.

Laws, F. & Schütze, H. Stopping criteria for active learning of named entity recognition Proceedings of Coling, 2008.

Raychaudhuri, S. Computational Text Analysis for Functional Genomics and Bioinformatics Oxford University Press, USA, 2006.

Roberts, A.; Gaizasukas, R.; Hepple, M. & Guo, Y. (ELRA), E. L. R. A. (ed.) Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008.

Settles, B. Collier, N.; Ruch, P. & Nazarenko, A. (ed.) Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets COLING 2004 (NLPBA/BioNLP) 2004, COLING, 2004, 107-110.

Vlachos, A. Evaluating and combining biomedical named entity recognition systems Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Association for Computational Linguistics, 2007.