Avaliação de técnicas de aprendizado de máquina para classificação de seções de laudos de biópsia renal auxiliada pela terminologia DeCS

Flávia P. Nicolas[1], K. J. Abraham[2], Amanda R. Reis[1], Ivan T. Pisa[1], Evandro E. S. Ruiz[3]

[1]Departamento de Informática em Saúde, UNIFESP, Rua Botucatu, 862, Vila Clementino, 04023-062 - São Paulo SP, Brasil

[2]Programa de Pós Graduação em Genética, FMRP – USP, Ribeirão Preto SP, Brasil

[3]Departamento de Computação e Matemática, FFCLRP - USP, Av. Bandeirantes, 3900, 14040-901 - Ribeirão Preto SP, Brasil

{flavia.nicolas, amanda.reis, ivan.pisa}@unifesp.br, abraham@iastate.edu, evandro@usp.br

Abstract. *Natural language processing is nowadays a valuable resource in biomedical informatics mainly due to the increasing demand for knowledge and the availability of documents in digital format. This paper describes the use and evaluation of five machine learning techniques applied to renal biopsy reports written as free text. The pre-processing and the named entity recognition tasks using the DeCS terminology is also reported.*

Resumo. *A grande demanda por conhecimentos novos e a crescente disponibilidade de documentos digitalizados fazem do processamento de linguagem natural um recurso indispensável ao tratamento da informação. Este artigo descreve a utilização e avaliação de cinco técnicas de aprendizado de máquina na classificação de seções de laudos de biópsia renal escritos em texto livre, bem como o pré-processamento dos documentos e o reconhecimento de termos médicos usando a terminologia DeCS.*

1. Introdução

Atualmente as pesquisas em Medicina e Biologia humana não prescindem de recursos de Processamento de Linguagem Natural (PLN) em suas várias tarefas. Nelas o objetivo último é a descoberta de novos conhecimentos a partir de relatos escritos. Como exemplo, os workshops de desafios da I2B2, como relatado por Uzuner e colegas [Uzuner 2010], descrevem bem a complexidade atual desses reptos de PLN no processamento de textos clínicos.

O objetivo deste trabalho é avaliar quais técnicas de aprendizado de máquina (AM) podem classificar seções de laudos de biópsia renal a partir de uma seleção de termos médicos reconhecidos em laudos escritos em texto livre. Para reconhecimento desses termos médicos, usamos a abordagem de dicionário. Neste caso, usamos a terminologia DeCS - Descritores em Ciências da Saúde[1] como base de termos.

[1] http://decs.bvs.br

O trabalho de Frunza O. e Inkpen [Frunza O. e Inkpen 2010] é o mais relevante para este artigo. Os autores estudam a indentificação e classificação de relações semânticas entre doenças e tratamentos em sentenças provenientes de sumários de artigos publicados. Neste trabalho eles usam, como atributos dos classificadores, termos da terminologia unificada UMLS - *Unified Medical Language System* [Bodenreider 2004]. Rink e colaboradores [Rink et al. 2010] trabalham numa abordagem semelhante usando um classificador SVM - *Support Vector Machine* para descobrir relações entre problemas médicos. O artigo recente de Jiang [Jiang 2011] demonstra uma das várias aplicações de técnicas de AM para extração de entidades nomeadas em textos clínicos. Lembramos que a extração de entidades nomeadas é ainda uma tarefa em progresso em PLN dada a precariedade de recursos terminológicos especializados em muitas línguas e também ao surgimento freqüente de novos termos médicos, como demontrado em [Patrick 2011]. O uso de abordagens baseadas em dicionários para extração de relações e entidades nomeadas também não é uma novidade, como podemos ver em [Patel e Cimino 2009].

2. Materiais e Métodos

Para realizar o presente trabalho foram utilizados 1002 laudos completos de biópsia renal, referente aos anos de 1994 a 2010, obtidos utilizando critérios de aleatoriedade a partir de uma base contendo 3728 laudos completos. Estes laudos correspondem a doenças renais diagnosticadas, provenientes de pacientes de todas as regiões do Brasil. Todos os laudos foram redigidos por um único patologista do Serviço de Patologia Renal do Hospital do Rim e Hipertensão (UNIFESP). Nenhum autor deste trabalho teve acesso aos nomes dos pacientes.

Os laudos são escritos em linguagem natural e estão estruturados em seções da seguinte maneira: Exame Macroscópico, Exame Microscópico, Imunofluorescência Direta, Diagnóstico e Observação, além de um cabeçalho com informações gerais. No entanto, ainda existe uma grande quantidade de laudos que não estão estruturados, ou seja, com as seções definidas. Para podermos aproveitar esses outros laudos precisamos implementar um classificador [Caffé, Perez e Baranauskas 2011].

Definimos que os classificadores usarão termos médicos como atributos. Com o objetivo de reconhecer os termos médicos presentes nos laudos, modelamos um reconhecedor de entidades nomeadas baseado em dicionário. Este reconhecedor atua em cada sentença do laudo procurando pelos termos completos. Como dicionário utilizamos a terminologia DeCS, extraída do UMLS [Bodenreider 2004].

Previamente a etapa de reconhecimento de entidades nomeadas, os laudos e o dicionário foram pré-processados. Inicialmente os acentos foram removidos e todas as palavras foram convertidas para letras minúsculas. Numa segunda etapa, foi efetuada a remoção de *stop-words*, ou seja, palavras muito frequentes, como artigos, preposições, pontuação, conjunções e pronomes. Por fim adotou-se a estratégia de *stemming* [Porter 2006] para a redução das formas flexionais das palavras dos laudos. Este pré-processamento visou minimizar o custo e otimizar os resultados do processamento.

Uma vez que os laudos e dicionário estavam devidamente pré-processados, foi calculada a freqüência de cada termo DeCS nos laudos de biópsia renal. Com estes dados, foi elaborado de maneira automática um arquivo no formato .arff, em que os

termos DeCS encontrados são os atributos para a classificação e as instâncias são a frequência dos termos em cada seção dos laudos. Este arquivo foi utilizado como entrada para cinco algoritmos de classificação, que foram escolhidos com base no trabalho de Frunza O. e Inkpen [Frunza O. e Inkpen 2010]. São estes os algoritmos de classificação: i) Árvores de Decisão J48 (modelo baseado em decisão); ii) Naïve Bayes e Complement Naïve Bayes - CNB (modelos probabilísticos); iii) AdaBoost (modelo de aprendizado adaptativo); e iv) ZeroR (modelo de única regra).

Todos os classificadores fazem parte da ferramenta Weka[2]. Os treinamentos e testes foram realizados utilizando a estratégia de *cross validation*, com 10 *folds* [Weiss e Indurkhya 1998].

3. Resultados parciais

Esta seção apresenta os resultados obtidos a partir do uso das técnicas de AM citadas anteriormente, em que o objetivo era utilizar termos DeCS previamente selecionados para classificar corretamente seções de laudos de biópsia renal escritas em texto livre.

A principal métrica de avaliação considerada foi *F-Measure*, que representa a média harmônica entre precisão e revocação, sendo precisão a porcentagem de seções classificadas corretamente, e revocação a porcentagem de seções identificadas como relevantes pelo classificador. Esta métrica é considerada adequada quando o conjunto de dados não está balanceado.

A Figura 1 mostra que o algoritmo de classificação que obteve o melhor desempenho para o problema em questão foi o modelo de árvores de Decisão J48, que, recursivamente, divide os exemplos em subconjuntos, tentando separar cada classe das demais, seguido pelos modelos probabilísticos CNB e Naive Bayes, que também apresentaram bons desempenhos (93,2%, 91,2% e 90%, respectivamente).

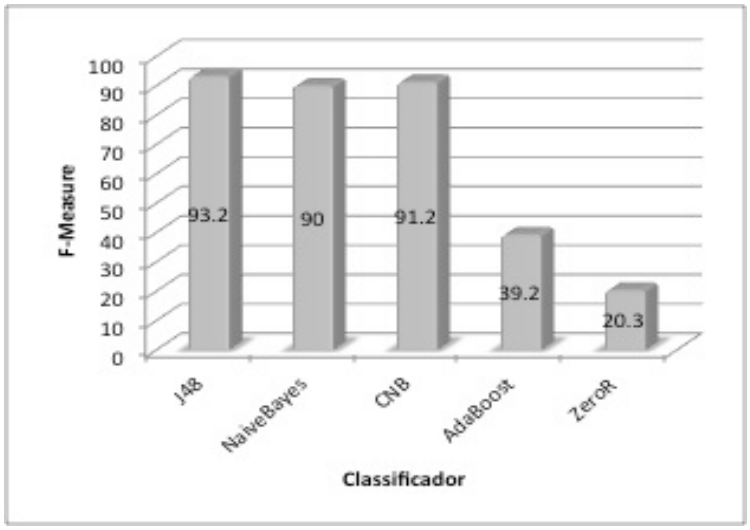

Figura 1. Resultado dos algoritmos de aprendizagem utilizados

O modelo de única regra ZeroR gerou o pior resultado, com apenas 20,3% de acertos, e o modelo de aprendizado adaptativo AdaBoost acertou apenas 39,2% dos

[2] http://www.cs.waikato.ac.nz/ml/weka/

casos. O desempenho do ZeroR pode ser explicado pelo fato de o algoritmo modelar uma base de dados com uma única regra. Este é o algoritmo de aprendizagem mais antigo do software Weka e, para uma base de dados, onde ocorrerá uma nova classificação, prediz o valor de maior frequência nos dados de treinamento [Witten e Frank, 1999].

Dados os resultados apresentados acima, pretendemos usar o classificador J48 para auxiliar na estruturação de outros laudos de biópsia renal. Ressaltamos que estes modelos de pré-processamento e de reconhecimento de entidades testados poderão ser úteis para nossa proposta maior de trabalhar com enriquecimento automático de ontologias.

4. Agradecimentos

Ao apoio financeiro do programa Professor Visitante do Exterior, da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil, concedido à K. J. Abraham e às contribuições dos professores Dr. José Augusto Baranauskas (DCM-USP) e Dr. Luiz Antônio Ribeiro de Moura, do Serviço de Patologia Renal do Hospital do Rim e Hipertensão (UNIFESP).

Referências

Uzuner Ö, South BR, Shen S, DuVall SL., 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text., J Am Med Inform Assoc. 2011 Sep-Oct;18(5):552-6.

Frunza O. e Inkpen D., Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences, Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 91–98, Uppsala, Sweden, 15 July 2010

Caffé Maria Izabela, Perez Pedro Santoro, Baranauskas José Augusto. Avaliação do Algoritmo de *Stacking* em Dados Biomédicos. In: XI Workshop de Informática Médica, Anais do XXXI Congresso da Sociedade Brasileira de Computação, Natal, RN, 19-22 de julho de 2011.

Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(suppl.1):D267–D270, January 2004.

Rink B, Harabagiu S, Roberts K., Automatic extraction of relations between medical concepts in clinical texts, J Am Med Inform Assoc. 2011 Sep-Oct;18(5):594-600.

Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, Xu H., A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, J Am Med Inform Assoc. 2011 Sep-Oct;18(5):601-6.

Patrick JD, Nguyen DH, Wang Y, Li M., A knowledge discovery and reuse pipeline for information extraction in clinical notes, J Am Med Inform Assoc. 2011 Sep-Oct;18(5):574-9.

Patel CO, Cimino JJ, Using semantic and structural properties of the Unified Medical Language System to discover potential terminological relationships, J Am Med Inform Assoc. 2009 May-Jun;16(3):346-53.

Porter MF. An algorithm for suffix stripping. Program: electronic library & information systems. 2006;40(3):211-218. http://dx.doi.org/10.1108/00330330610681286.

Weiss SM, Indurkhya N, Predictive Data Mining: A Practical Guide, MK, San Francisco, CA, 1998.

Witten I. H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco, 1999