Um Processo Baseado em Parágrafos para a Extração de Tratamentos em Artigos Científicos do Domínio Biomédico

Juliana Lilian Duque[1], Pablo Freire Matos[1], Cristina Dutra de Aguiar Ciferri[2], Thiago Alexandre Salgueiro Pardo[2], Ricardo Rodrigues Ciferri[1]

[1]Departamento de Computação – Universidade Federal de São Carlos
Rodovia Washington Luís, km 235 – 13565-905 – São Carlos – SP – Brasil

[2]Departamento de Ciências de Computação – Universidade de São Paulo
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brasil

{juliana_duque, pablo_matos, ricardo}@dc.ufscar.br,
{cdac, taspardo}@icmc.usp.br

Abstract. *This paper addresses the problem of extracting information from unstructured documents written in English and stored in PDF format. We propose a process that uses machine learning, rules and dictionary to identify and extract treatments in biomedical domain papers. We argue that searching firstly for sentences that contain complications can improve the efficiency of the identification and extraction of treatments, since the treatments mainly occurs in sentences with complications or in sentences very near in the same paragraph. The experiments showed that the proposed process obtained 88% of precision in the classification and 70% of recall in the extraction of treatments based only on a set of rules developed with Part-Of-Speech.*

Resumo. *Este artigo investiga o problema de extrair informações relevantes em documentos não estruturados no formato PDF escritos em inglês. Nós propomos um processo que usa aprendizado de máquina, regras e dicionário para identificar e extrair tratamentos em artigos do domínio biomédico. A busca inicial de sentenças que possuem complicações melhora a eficiência na identificação e extração de termos de tratamentos. Isso acontece porque tratamentos ocorrem principalmente na mesma sentença de complicação ou em sentenças próximas no mesmo parágrafo. Os experimentos mostraram uma precisão de 88% na classificação das sentenças e uma revocação de 70% na extração de tratamentos usando regras baseadas em Part-Of-Speech.*

1. Introdução

Atualmente na área médica uma grande quantidade de dados tem sido produzida e armazenada em documentos no formato textual não estruturados, conduzindo para a criação de volumosos conjuntos de documentos digitais [Stavrianou et al. 2007]. O volume de dados armazenados ultrapassa em muito as habilidades humanas de interpretá-los individualmente, exigindo técnicas para automatizar e analisar os documentos de forma ágil e preferencialmente de forma automática ou semiautomática. Devido à alta taxa de crescimento de documentos textuais, a qual é medida em termos do número de publicações de artigos e revistas, torna-se impossível analisar toda a

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 124–133,
Cuiabá, MT, Brazil, October 24–26, 2011.

literatura médica relevante manualmente, mesmo em tópicos específicos [Jensen et al. 2006].

O surgimento da Mineração de Textos (MT) foi motivado pela necessidade de se descobrir de forma semiautomática informações e conhecimento novos em textos. O uso das ferramentas de MT torna-se indispensável neste cenário, possibilitando o processamento de uma grande quantidade de textos, permitindo recuperar informações relevantes, possibilitando a extração de informação automática e o reconhecimento de padrões [Ebecken et al. 2003; Gupta and Lehal 2009].

Este artigo utiliza a MT para identificar e extrair informações úteis, novas e interessantes em artigos científicos do domínio biomédico, os quais estão no formato PDF e escritos em inglês, mais especificamente aplicada no estudo de caso em artigos completos da doença Anemia Falciforme. A Anemia Falciforme (AF) ou *Sickle Cell Anemia* (SCA) é uma doença hematológica e hereditária, que causa a destruição crônica das células vermelhas do sangue, afetando principalmente a população negra e considerada como um problema de saúde pública no Brasil [Pinto et al. 2009].

A extração de informação enfocará especificamente em termos de "tratamentos" (i.e. drogas, terapias e procedimentos usados para tratar uma doença). Para alcançar esse objetivo é proposto um processo que combina três abordagens para a extração de informação: aprendizado de máquina, regras e dicionário. A técnica de aprendizado de máquina é utilizada exclusivamente na classificação de sentenças, visando filtrar o conjunto de sentenças para um subconjunto de sentenças de interesse, enquanto as regras e o dicionário são usadas para a identificação e a posterior extração de termos de tratamentos nas sentenças de interesse. Considera-se como hipótese deste trabalho que na maioria dos casos os termos de tratamentos ocorrem na mesma sentença de uma complicação ou em sentenças próximas em um mesmo parágrafo. Portanto, o parágrafo é considerado neste processo como uma unidade com conteúdo de informações centralizado, no qual se localiza a informação de interesse (i.e. termos de tratamentos). Esta hipótese foi baseada em uma análise empírica de artigos da doença Anemia Falciforme, para os quais descobriu-se que: (i) ~10% dos tratamentos apareceram na sentença anterior ou posterior à sentença na qual a complicação foi encontrada; (ii) ~90% dos tratamentos apareceram na mesma sentença na qual a complicação foi encontrada; e (iii) muitos poucos tratamentos ocorreram em sentenças mais distantes.

Este artigo está estruturado da seguinte forma. Na Seção 2 são descritas as principais abordagens encontradas na literatura para extrair informação; na Seção 3, é proposto o processo para a identificação e a extração de termos de tratamentos, enquanto na Seção 4 é realizado um estudo de caso no qual o processo é aplicado e a sua eficiência medida; na Seção 5 são resumidos os trabalhos correlatos e, na Seção 6, são apresentadas as conclusões e as propostas de trabalhos futuros.

2. Abordagens para Extração de Informação

Kou et al. (2005) e Cohen and Hunter (2008) descrevem duas abordagens para a extração de informação: abordagem baseada em regras, utilizada para identificar padrões de extração com o uso de expressões regulares; e abordagem baseada em aprendizado de máquina, que utiliza classificadores para separar ou identificar sentenças de interesse. Além dessas, Krauthammer and Nenadic (2004) apresentam uma terceira abordagem para o reconhecimento automático de termos: abordagem baseada em

dicionário, que utiliza informações para auxiliar no reconhecimento dos termos ou das entidades no texto.

A abordagem **baseada em aprendizado de máquina** utiliza classificadores para separar ou identificar sentenças de interesse [Cohen and Hunter 2008]. Técnicas de aprendizado de máquina são utilizadas em reconhecimento automático de termo, que são projetadas para atender a uma classe específica de entidades e usam dados de treinamento para aprender as características que são úteis e relevantes para o reconhecimento e a classificação de termos [Krauthammer and Nenadic 2004]. Os principais problemas relacionados aos algoritmos de aprendizado de máquina são a necessidade de grandes quantidades de dados de treinamento e o fato que a classificação é prejudicada quando o conjunto de dados de uma classe é pequeno em relação a outras classes [Ananiadou and McNaught 2006].

A **abordagem baseada em regras** é utilizada para identificar padrões de extração com expressões regulares. Esta abordagem é normalmente difícil de se ajustar a diferentes domínios ou classes, uma vez que as regras são específicas do domínio [Ananiadou and McNaught 2006]. Outra desvantagem dessa abordagem é o tempo significativo para a definição e para a validação das regras [Cohen and Hunter 2008].

A técnica de **Processamento de Língua Natural *Part-Of-Speech*** (POS) pode ser utilizada para auxiliar as expressões regulares na identificação dos termos relevantes contidos em uma sentença. O etiquetador POS consiste em rotular as palavras segundo a sua classe gramatical. Substantivo, adjetivo, advérbio, verbo e preposição são alguns exemplos de classes gramaticais [Matos 2010]. Regras simples sem POS é um tipo de regra convencional, em que o desenvolvimento do padrão é dependente do domínio; regras apenas com POS utilizam padrões POS mais específicos visando extrair termos com baixa ocorrência de falsos positivos; e regras com POS e termos representativos utilizam-se de termos representativos (e.g. verbo ou palavra) para identificar se uma sentença contém ou não um termo [Matos 2010].

A **abordagem baseada em dicionário** dispõe de uma lista de termos para localizar as ocorrências no texto. Considera-se um termo ocorrência de cada sequência de palavras no texto que corresponder a uma entrada no recurso terminológico; apenas cadeias de caracteres são tratadas como tais termos [Ananiadou and McNaught 2006]. Uma desvantagem desta abordagem é a restrição de nomes que estão presentes no dicionário e o falso reconhecimento causado principalmente por nomes curtos e baixa revocação devido a variações de ortografia [Tsuruoka and Tsujii 2004].

3. Processo para Extração de Tratamento

Na Figura 1 são apresentadas as etapas do processo proposto para efetuar a extração de informação sobre "tratamentos" em artigos científicos do domínio biomédico. Este processo é composto de **cinco etapas**. Essencialmente, utiliza-se de dois classificadores, chamados de C1 (Classificador 1) e de C2 (Classificador 2), ambos com o objetivo de separar as sentenças de interesse que provavelmente terão, respectivamente, termos de complicação e termos de tratamento, das sentenças que possivelmente não terão nenhum destes termos. Usa-se também dicionários e regras para identificar e extrair as partes de interesse dentro das sentenças pré-selecionadas. A nossa hipótese é que as sentenças que possuem termos de tratamento ocorrem em uma sentença que possui um termo de complicação ou ocorre em sentenças próximas em um mesmo parágrafo. Desta forma,

primeiramente procura-se por sentenças com termos de complicação para posteriormente identificar sentenças que provavelmente possuirão termos de tratamento. Uma visão geral do processo é apresentada na Figura 1.

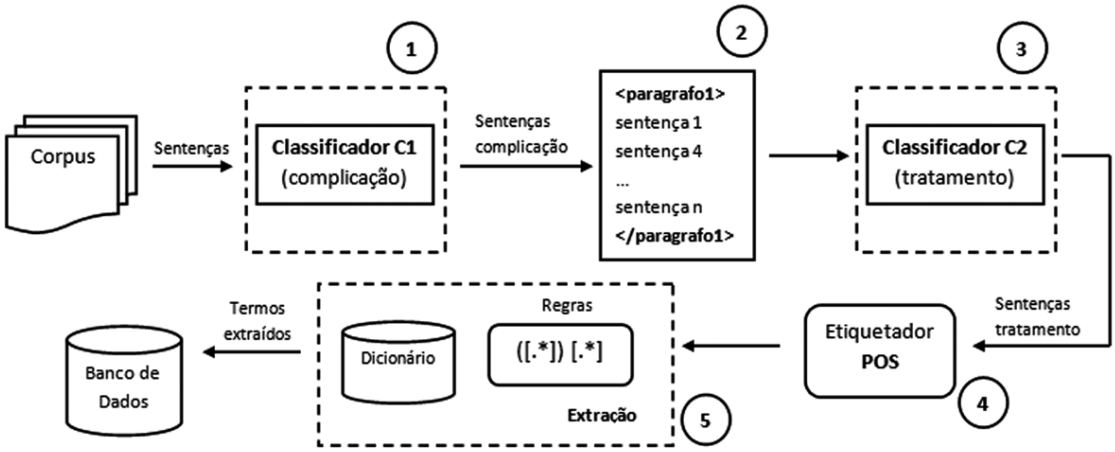

Figura 1. Processo para extração de tratamento

O Classificador C1 tem como objetivo identificar as sentenças que possivelmente possuem termos de complicação e o Classificador C2 identificar as sentenças que possivelmente possuem termos de tratamento. As sentenças direcionadas aos classificadores estão divididas em duas classes de interesse: "complicação" e "outros" para o Classificador C1 e; "tratamento" e "outros" para o Classificador C2.

Com o intuito de encontrar termos de tratamento, primeiramente as sentenças são classificadas em sentenças com termos de complicação utilizando o Classificador C1 (i.e. passo 1 da Figura 1). As sentenças de interesse são agrupadas adicionando-se as sentenças que participam do mesmo parágrafo (passo 2 da Figura 1) e estas são enviadas ao Classificador C2 (passo 3 da Figura 1). As sentenças identificadas no passo 3 como possivelmente tendo termos de tratamento são etiquetadas conforme sua classe gramatical (passo 4 da Figura 1). Substantivo, adjetivo, advérbio, verbo e preposição são alguns exemplos de classes gramaticais. Após este etapa, é realizado o processo de extração dos termos de tratamentos (passo 5 da Figura 1), utilizando as abordagens de dicionário e regras. O dicionário, o qual deve ser criado por um especialista do domínio, é usado na pesquisa exata dos termos de tratamentos armazenados comparando-os com os termos relevantes existentes ou não nas sentenças. No processo proposto para a extração de termos de tratamentos, utiliza-se adicionalmente um dicionário com variações de termos e com sinônimos, de forma a reduzir os problemas da técnica de dicionário e com isto melhorar a precisão da extração de termos conhecidos de tratamentos. Já o uso de regras permite encontrar novos tratamentos desconhecidos que venham a surgir na literatura. Ao final do processo, é efetuado a aplicação do conjunto de regras nas sentenças de tratamentos e os termos relevantes e interessantes extraídos são armazenados em um banco de dados (passo 6 da Figura 1). Cada um desses passos é exemplificado nas próximas seções.

4. Estudo de Caso

O estudo de caso investigou separadamente a eficiência na classificação das sentenças de interesse na fase de filtragem das sentenças e da eficiência da extração de informação

de novos tratamentos usando regras com POS e também com POS combinado com termos representativos.

4.1. Classificação de Sentenças

Inicialmente foi realizada uma classificação manual em um conjunto de 765 sentenças existente no corpus. Essas sentenças foram examinadas de duas formas: 1) 765 sentenças analisadas e compreendidas como sendo sentenças de "complicação" e "outros" e; 2) as mesmas 765 sentenças analisadas e mencionadas como sendo sentenças de "tratamento" e "outros". Para a primeira representação, esse cenário foi utilizado para preparar o desenvolvimento do Classificador C1 e a segunda para o Classificador C2. Todas as sentenças foram utilizadas para o treinamento e o teste dos classificadores usando o método de particionamento *10-fold cross validation*. A porcentagem da distribuição das classes pode ser vista na Tabela 1:

Tabela 1. Porcentagem da distribuição das classes

Classificador	Quantidade	Sentenças	Porcentual
C1	**337**	**Complicação**	**44%**
	428	**Outros**	**56%**
C2	**394**	**Tratamento**	**51%**
	371	**Outros**	**49%**

O experimento é descrito como segue: primeiramente foi realizada uma limpeza no conjunto de treinamento, removendo pontuação, parênteses, colchetes e apóstrofes. Após, a matriz de atributo-valor foi construída usando frequência mínima de dois para a seleção de atributo. A seleção de atributo foi composta de 1 a 3 gramas. Não foram usadas a técnica de *stemmer* ou eliminação de *stopwords*. O uso de *stopwords* deve-se ao fato de que alguns termos ajudam na formação de regras para a extração dos termos.

Dois algoritmos de aprendizado de máquina foram utilizados para o experimento: *Support Vector Machine* (SVM) e *Naive Bayes*. Também foram utilizadas quatro configurações de pré-processamento gerando 8 configurações. Os filtros utilizados foram: 1) *No Filter*; 2) *Randomize*; 3) *Remove Misclassified* para remover ruído; e 4) *Resample*, método utilizado para balancear as classes de interesse.

Os testes foram executados usando o classificador SCA-Classifier do ambiente do Projeto SCA [Matos 2010], o qual foi desenvolvido na linguagem de programação Java com o uso da API do ambiente Weka [Witten and Frank 2005]. O SVM apresentou o melhor resultado para os classificadores C1 e C2, conforme observado na Figura 2 (a) e (b), respectivamente. Assim, o Classificador C1 (Figura 2 (a)) e o Classificador C2 (Figura 2 (b)) foram construídos a partir do melhor resultado apresentado.

Na Tabela 2 é apresentado o resultado da classificação automática das 359 novas sentenças usadas para os dois classificadores. Foram utilizadas as principais métricas usadas em sistemas de extração de informação como precisão, revocação e medida-F e em sistemas de aprendizado de máquina como acurácia [Matos 2010]. Os resultados mostram que as classificações de sentenças com complicação e com tratamento obtiveram uma boa precisão compatível com valores de precisão obtidos em outros trabalhos que realizam extração de informação de artigos completos. Apesar da revocação não ter sido alta, a repetição de termos de tratamentos nas sentenças faz com que a perda de algumas sentenças não tenha tanto impacto no processo de extração de informação, conforme mostrado no próximo experimento.

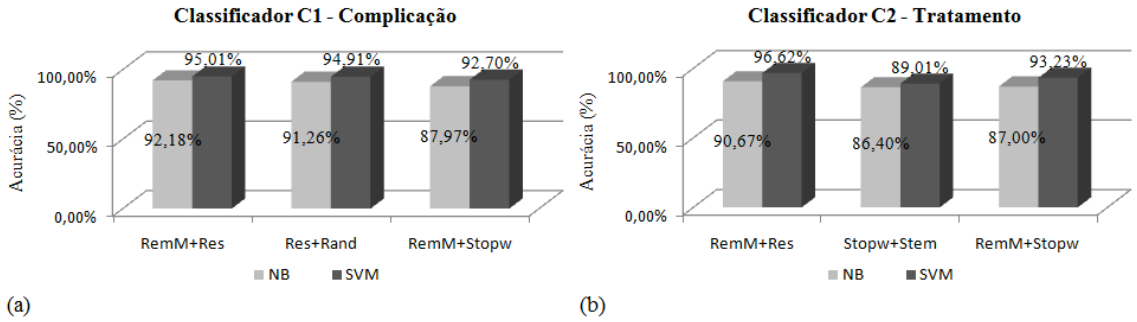

Figura 2. Melhores resultados para os classificadores C1 (a) e C2 (b)

Tabela 2. Resultado da classificação automática

Classificador	Qtde	Sentenças	Precisão	Revocação	Acurácia	Medida-F
C1	120	Complicação	85%	64%	79%	73%
C2	107	Tratamento	88%	51%	71%	64,5%

4.2. Regras

Para descobrir padrões manualmente no corpus, inicialmente foi necessário analisar um subconjunto de sentenças nas quais continham um requisito principal: termos de tratamentos ou palavras chaves que indicassem tratamentos, tais como: *study, trial, experiment*. Este subconjunto foi reaproveitado do conjunto de sentenças de tratamentos utilizados para treinamento do Classificador C2, para o qual foram selecionadas 394 sentenças. Ainda, foram removidas as sentenças que não continham palavras chaves que indicassem ou que continham o termo relevante. Por decorrência desta filtragem, remanesceram 123 sentenças, nas quais continham exatamente os termos de tratamento relevantes, termos estes que já haviam sido avaliados e considerados pelo especialista da área, no caso um especialista da doença Anemia Falciforme.

O subconjunto resultante de 123 sentenças totalizou 163 termos de tratamentos relevantes. Alguns termos estavam presentes em várias sentenças, enquanto alguns termos de tratamentos foram encontrados numa mesma sentença mais que uma vez, tais como termos de *hydroxyurea* e *transfusion*. Já outros termos foram encontrados apenas uma vez, tais como: *bone marrow transplantation* e *penicillin*. Logo, o subconjunto resultante de sentenças foi processado, etiquetado pelo padrão POS, e ainda, para facilitar, foi agrupado por grau de semelhança.

Após um processo manual, foram descobertos padrões e gerado um conjunto de 13 regras, subdividido em duas estratégias: 1) Verbo ou palavra representativa com POS e 2) Somente POS. Entende-se por verbo ou palavra representativa uma informação que pode identificar um termo relevante na sentença. O motivo de criar duas estratégias é que para a primeira, o processamento é realizado em parte da sentença, ou seja, a expressão regular somente casará com a sentença se, e somente se, existir o verbo ou palavra representativa na sentença. Caso o verbo ou palavra representativa existir, padrões POS são utilizados para extrair o termo relevante somente em uma parte específica da sentença. A vantagem de extrair informação em somente uma parte específica da sentença é que com isso diminui-se a possibilidade de se extrair falsos

positivos [Matos 2010]. Para o segundo tipo de regra, o uso somente de POS, o processamento é realizado na sentença por completo, fazendo com que o padrão POS case com algum padrão POS descoberto e criado através da análise realizada previamente no subconjunto de sentenças. Na Tabela 3 e Tabela 4 são demonstrados exemplos para as duas estratégias, respectivamente. *Treatment* e *therapy* são termos representativos presentes na regra 1. Para entendimento, o caractere \w significa uma sequência de letras, números ou sublinhado, a etiqueta IN indica preposição, NN indica substantivo no singular, NNP indica nome próprio, NNS indica substantivo comum no plural e as etiquetas VBD e VBN indicam verbos.

Tabela 3. Exemplo de regra da Estratégia 1

[\w-/\\]*_IN (?:[\w-/\\]*)?([\w-/\\]*_NN|[\w-/\\]*_NNP|[\w-/\\]*_NNS) (?:treatment_NN|therapy_NN)

Tabela 4. Exemplo de regra da Estratégia 2

(?:[\w-/\\]*_NNS) (?:[\w-/\\]*_IN|[\w-/\\]*_VBD|[\w-/\\]*_VBN)* ([\w-/\\]*_NN|[\w-/\\]*_NNP|[\w-/\\]*_NNS) (?:[\w-/\\]*_NN)?

Após a geração do conjunto de regras, as regras foram aplicadas no subconjunto de sentenças para obter o cálculo da precisão e revocação. Dos 163 termos de tratamentos contidos no subconjunto de 123 sentenças, 114 termos foram identificados corretamente, 139 falsos positivos e 49 termos não foram identificados pelo conjunto de regras. A precisão do conjunto de regras resultou em 45%, e a revocação resultou em 70%. Estes resultados foram obtidos para a extração de novos tratamentos desconhecidos, e para os tratamentos conhecidos, o processo proposto utiliza-se de um dicionário estendido com variações e siglas visando melhorar a eficiência da extração de tratamentos nas sentenças.

4.3. Dicionário

Nomeado como Dicionário, o banco de dados biomédico contém termos de tratamentos relevantes avaliados pelo especialista da área. Conforme já citado, usamos uma técnica alternativa para reduzir os problemas de restrição de nomes em abordagens baseada em dicionário. Para tanto, com o uso de variações de termos e siglas, foi possível identificar 100% das ocorrências de tratamentos já conhecidas. Exemplos de termos armazenados no dicionário são: *hydroxyurea* e sua variação *HU*, *placebo* e *antibiotic*.

5. Trabalhos Correlatos

Na literatura biomédica encontram-se trabalhos correlatos que extraem informação de resumos ou artigos completos. A maioria destes trabalhos extrai informação do MEDLINE, são baseados em entidades de genes e proteínas e usam as três abordagens de extração de informação apresentadas na Seção 2.

Dos trabalhos correlatos apresentados, Yang et al. (2009) e Tanabe and Wilbur, (2002a,b) têm como objetivo extrair informação; e Bremer et al. (2004) e Matos et al. (2010) possuem o propósito de povoar um banco de dados. Yang et al. (2009) extraem informação de resumos, enquanto Bremer et al. (2004), Tanabe and Wilbur (2002a,b) e Matos et al. (2010) extraem informação de artigos completos. Yang et al. (2009) empregam POS e utilizam abordagem de regras para extrair informação de resumos. Destacam-se trabalhos correlatos que extraem informação de artigos completos: Bremer

et al. (2004) utilizam regras e dicionário como abordagens de extração de informação e não fazem uso de POS em seu conjunto de regras; Tanabe and Wilbur (2002a,b) e Matos et al. (2010), ambos utilizam a combinação das três abordagens: aprendizado de máquina, regras e dicionário, e ainda, utilizam etiquetadores POS. Tanabe and Wilbur (2002a,b) apresentaram o maior percentual de precisão (72,5%) e revocação (50,7%) em extração de artigos completos.

Embora o trabalho de Matos et al. (2010) também utilize a abordagem de aprendizado de máquina para classificar sentenças de interesse e de regras e de dicionários para identificar e extrair informação, o processo proposto neste artigo inova no sentido de usar um *pipeline* de classificação de sentenças com a extensão das sentenças de interesse em nível de parágrafo antes da segunda classificação. Desta forma, primeiro classifica-se sentenças que provavelmente possuirão termos de complicação, faz-se a expansão das sentenças considerando sentenças próximas as sentenças classificadas em nível de parágrafo e depois classifica-se as sentenças que provavelmente possuirão termos de tratamento. Ademais, não é possível usar nem o dicionário nem as regras propostas no trabalho de Matos et al. (2010) desde que são específicas para termos de efeitos positivos e negativos, enquanto este artigo enfoca em termos de tratamento. As mesmas considerações aplicam-se aos demais trabalhos correlatos que usam regras e dicionários.

6. Conclusão e Trabalho Futuro

Este artigo propôs um processo para extrair termos de tratamentos em artigos científicos do domínio biomédico no formato PDF e escritos em inglês. O processo utiliza um *pipeline* de classificação para selecionar sentenças de interesse. A nossa hipótese é que termos de tratamento geralmente ocorrem na mesma sentença que possui termos de complicação ou ocorre em sentenças próximas em um mesmo parágrafo. Assim, após uma primeira classificação das sentenças que provavelmente possuem termos de complicação, as sentenças são expandidas de forma a englobar também sentenças próximas em um mesmo parágrafo. Estas sentenças expandidas são então classificadas para filtrar sentenças com termos de tratamento.

Para validar o processo proposto, um estudo de caso foi realizado para analisar a qualidade dos classificadores e a qualidade da extração de tratamentos usando regras POS para descobrir novos tratamentos. Os resultados mostraram que a classificação obteve uma boa precisão e apesar da revocação não ser muito alta, isto não impactou na fase seguinte de extração de informação desde que os termos de tratamentos se repetem ao longo do texto completo de um artigo. Vale destacar que com o processo proposto o tempo de processamento de cada artigo pode ser realizado rapidamente em um computador desktop típico, com tempo próximo a 1,5 minuto.

Como extensão desta pesquisa espera-se adaptar e aplicar o processo proposto para outros termos no contexto da doença Anemia Falciforme. O uso de índices para agilizar a identificação de termos também será investigado.

Referências

Ananiadou, S.; McNaught, J. (2006) (Ed.). Text mining for biology and biomedicine. Norwood, MA: Artech House, 302 p.

Bremer, E. G. et al. (2004) Text mining of full text articles and creation of a knowledge base for analysis of microarray data. In: KELSI, 2004, Milan, Italy. Proceedings. 2004. p. 84-95.

Cohen, K. B.; Hunter, L. (2008) Getting started in text mining. PLoS Computational Biology, v. 4, n. 1, p. 1-3.

Ebecken, N. F. F.; Lopes, M. C. S.; Costa, M. C. D. A. (2003) Mineração de textos. In: Rezende, S. O. (Eds.). Sistemas inteligentes: fundamentos e aplicações. São Carlos: Manole, p. 337-370. cap. 13.

Gupta, V.; Lehal, G. (2009) A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence, v. 1, n. 1, p. 60-76.

Jensen, L. J.; Saric, J.; Bork, P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics, v. 7, n. 2, p. 119-129.

Kou, Z.; Cohen, W. W.; Murphy, R. F. (2005) High-recall protein entity recognition using a dictionary. Bioinformatics, v. 21, p. i266-273. Suppl. 1.

Krauthammer, M.; Nenadic, G. (2004) Term identification in the biomedical literature. Journal of Biomedical Informatics, v. 37, n. 6, p. 512-526.

Matos, P. F. (2010) Metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico. 159 f. Dissertação (Mestrado em Ciência de Computação) – Departamento de Computação, Universidade Federal de São Carlos, São Carlos.

Matos, P. F. et al. (2010) An environment for data analysis in biomedical domain: information extraction for decision support systems. In: García-Pedrajas (Eds.). IEA-AIE. 23th. Heidelberg: Springer, p. 306-316.

Pinto, A.C.S. et al. (2009) Technical Report Sickle Cell Anemia. Technical Report, Federal University of São Carlos, http://sca.dc.ufscar.br/download/files/report.sca.pdf

Stavrianou, A.; Andritsos, p.; Nicoloyannis, N. (2007) Overview and semantic issues of text mining. Sigmod Rec., v. 36, n. 3, p. 23-34.

Tanabe, L.; Wilbur, W. J. (2002) Tagging gene and protein names in biomedical text. Bioinformatics, v. 18, n. 8, p. 1124-1132, 2002a.

_____. Tagging gene and protein names in full text articles. (2002) In: WNLPBD, Phildadelphia, Pennsylvania. Proceedings. Morristown, NJ: Association for Computational Linguistics, 2002b. p. 9-13.

Tsuruoka, Y.; Tsujiii, J. I. (2004) Improving the performance of dictionary-based approaches in protein name recognition. Journal of Biomedical Informatics, v. 37, n. 6, p. 461-470.

Witten, I. H.; Frank, E. (2005) Data mining: pratical machine learning tools and techniques with Java implementations. 2nd ed. San Francisco, CA: Morgan Kaufmann, 525 p.

Yang, Z.; Lin, H.; Wu, B. (2009) BioPPIExtractor: A protein-protein interaction extraction system for biomedical literature. Expert Syst. Appl., v. 36, n. 2, p. 2228-2233, 2009.