Avaliação de um método de mapeamento de laudos médicos para uma representação estruturada: estudo de caso com laudos de endoscopia digestiva alta

Daniel de Faveri Honorato[1], Maria Carolina Monard[2], Huei Diana Lee[1], Antonio Pietrobom Neto[3], Wu Feng Chung[1]

[1]Laboratório de Bioinformática
Universidade Estadual do Oeste do Paraná (UNIOESTE)
Parque Tecnológico Itaipu (PTI)
Foz do Iguaçu - Paraná - Brasil CEP 85856-970

[2]Departamento de Ciência da Computação
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos - São Paulo - Brasil CEP 13560-970

[3]Serviço de Endoscopia Digestiva
Hospital Municipal de Paulínia
Paulínia - São Paulo - Brasil

dfaverih@gmail.com, mcmonard@icmc.usp.br, hueidianalee@gmail.com, pietrobom@mpcnet.com.br, wufengchung@gmail.com

Abstract. *In order to perform the Text Mining process, text data should be pre-processed so that predictive or descriptive methods are applicable. To this end, we have proposed a method that transforms information related to medical findings described in natural language to the attribute-value format. Two modes can be used to apply the proposed method: automatic, where only syntactical information is used, and semi-automatic where semantic information can be added by domain specialists. This work shows a case study where this method was applied using the automatic mode. Although applying the method using this mode represents its worst case as it ignores any semantic information, results show that our proposal is feasible even in its worst case.*

Resumo. *O processo de Mineração de Textos pode auxiliar especialistas na tomada de decisão por meio de extração de padrões a partir de tabelas atributo-valor. Neste trabalho são apresentados os resultados experimentais da aplicação de um método proposto anteriormente o qual pode ser aplicado automaticamente ou semi-automaticamente com a ajuda de especialistas do domínio, e tem por objetivo realizar o mapeamento de documentos não estruturados para uma tabela atributo-valor. Foi analisada uma coleção de 6000 laudos de Endoscopia Digestiva Alta, com bons resultados alcançados, mostrando que o método pode auxiliar na redução do tempo de atuação dos especialistas na análise de grandes quantidades de documentos não estruturados.*

1. Introdução

Com o enorme crescimento das bases de dados, resultado do avanço tecnológico ocorrido nos últimos anos, tornou-se cada vez mais difícil analisar e extrair, manualmente, informações e padrões a partir dos dados. Na área médica esse crescimento é perceptível, uma vez que uma quantidade considerável das informações de pacientes e processos laboratoriais estão descritas em laudos e formulários médicos e, muitas dessas informações encontram-se atualmente armazenadas eletronicamente. Assim, essas informações podem ser analisadas em busca de padrões que auxiliem, por exemplo, no processo de tomada de decisão. No entanto, a análise manual de um conjunto grande de informações desse tipo é inviável, pois trata-se de uma tarefa que tem alto custo de tempo e que está sujeita à subjetividade [Honorato et al. 2008a, Lee 2005]. Essas informações podem estar representadas em diferentes formatos, sendo que um dos formatos comumente utilizados é o formato textual não estruturado. Para que esses dados textuais brutos possam tornar-se úteis, é necessário que eles sejam representados de maneira apropriada para a extração de padrões, tal que um modelo que represente o conhecimento embutido nessas informações possa ser construído. Uma das maneiras de alcançar esse objetivo é por meio do processo de Mineração de Textos [Feldman and Sanger 2006]. Esse processo consiste, basicamente, em três fases principais: (1) pré-processamento dos textos; (2) extração de padrões; e (3) pós-processamento. Na fase de pré-processamento, o conjunto de textos é transformado para uma representação adequada para ser utilizada pelos algoritmos de extração de padrões. Geralmente, os documentos são representados em uma tabela atributo-valor, na qual palavras selecionadas do conjunto de documentos são transformadas em atributos. Essa tabela é então utilizada na fase de extração de padrões para construir modelos. Nessa etapa podem ser utilizados, por exemplo, algoritmos de inteligência artificial da área de aprendizado de máquina. Os modelos induzidos podem ser representados por estruturas simbólicas como árvores de decisão e regras de produção, as quais permitem maior compreensibilidade humana [Witten and Frank 2005]. Esses modelos construídos são analisados e validados na fase de pós-processamento.

Em [Honorato 2008, Honorato et al. 2008b] foi proposto e implementado um método geral para o pré-processamento de documentos não estruturados, ou seções específicas desses documentos, que tem como resultado a representação em uma tabela atributo-valor das informações contidas nesses documentos. Esse método pode ser aplicado a qualquer conjunto de documentos textuais que verifiquem as seguintes duas propriedades:

1. as informações são descritas utilizando um vocabulário controlado; e
2. as informações consistem de frases assertivas simples.

Deve ser observado que laudos médicos verificam ambas propriedades. Um dos principais objetivos desse método é diminuir, sempre que possível, a intervenção dos especialistas, fornecendo, nas diversas fases do método, informações que facilitam o trabalho a ser realizado pelos especialistas[1], que são os responsáveis pela inclusão do aspecto semântico das informações. O método pode ser aplicado, de modo automático, sem intervenção dos especialistas, considerando somente o aspecto morfo-sintático das

[1]Em [Honorato et al. 2008a] foi proposto outro método para realizar essa tarefa, mas ele requer uma intensa interação com os especialistas do domínio, limitando a sua aplicação, o que motivou o desenvolvimento deste novo método.

informações, *i.e.*, sem levar em conta o aspecto semântico. No entanto, ele pode também ser aplicado de modo não automático com a participação de especialistas do domínio. Nesse caso, o especialista é responsável por levar em conta o aspecto semântico das informações que estão sendo processadas. Assim, caso houver participação do especialista, informações específicas do domínio podem ser utilizadas e melhores resultados serão alcançados. O ambiente computacional *Term Pattern Discover* — TP-DISCOVER — implementa todas as fases do método desenvolvido [Honorato and Monard 2008].

Este trabalho tem por objetivo apresentar os resultados experimentais da aplicação do método proposto em [Honorato 2008] sobre um conjunto de laudos de Endoscopia Digestiva Alta — EDA —, especificamente, informações relacionadas ao estômago.

O trabalho está organizado da seguinte maneira: na Seção 2 é descrito resumidamente o método desenvolvido; na Seção 3 é apresentado o formato do laudo utilizado neste trabalho; na Seção 4 é apresentada a avaliação experimental realizada utilizando variações dos valores dos diferentes parâmetros do método e, ao final, na Seção 5 são apresentadas as considerações finais.

2. Método

O método proposto [Honorato 2008], o qual foi implementado no ambiente computacional TP-DISCOVER, é composto por cinco fases: (1) Pré-processamento; (2) Extração de terminologia —ET; (3) Identificação de atributos; (4) Construção do dicionário; e (5) Construção da tabela atributo-valor. Essas fases, descritas a seguir, estão ilustradas na Figura 1.

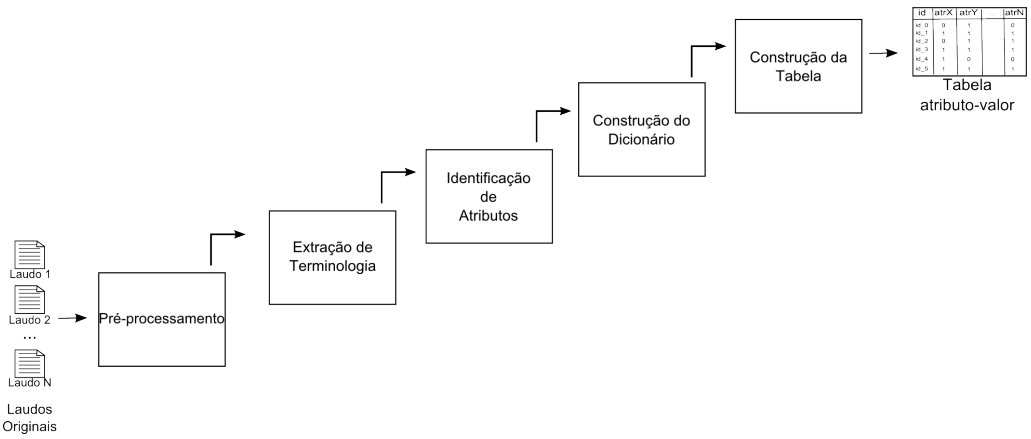

Figura 1. Método desenvolvido

Pré-processamento – Essa fase consiste de sete etapas, as quais são aplicadas sobre o conjunto de laudos do domínio que está sendo trabalhado. Essas etapas correspondem à preparação do *corpus*[2] por meio de tarefas como: divisão do documento de acordo com as seções que ele contém; construção do conjunto de frases únicas que consiste de todas as frases diferentes do conjunto de documentos; remoção de *stopwords* que são palavras consideradas não relevantes para a análise do texto; transformação para minúsculo; correção ortográfica; aplicação de substituições tais como sinônimos ou frases que mapeiam mais

[2]Em lingüística, um *corpus* consiste em um conjunto de textos, os quais são utilizados em análises estatísticas, verificação de ocorrências e validação de regras lingüísticas em um universo específico.

de um evento; e aplicação do lematizador, o qual transforma verbos para a forma infinitiva e substantivos e adjetivos para masculino singular.

Extração de Terminologia – O objetivo desta fase é determinar as palavras mais apropriadas para serem consideradas como unidades terminológicas, as quais serão utilizadas na próxima fase. Para a identificação das unidades terminológicas é adotada uma abordagem híbrida utilizando tanto conhecimento estatístico quanto linguístico, seguida da aplicação de algumas heurísticas sobre o conjunto de termos identificados. Na abordagem híbrida implementada no método, após uma análise morfo-sintática prévia dos termos do domínio são geradas duas listas, uma lista de unigramas contendo palavras que casam com a classe gramatical N (*substantivo*) e outra lista de bigramas contendo palavras que casam com palavras consecutivas da classe gramatical N N. Primeiramente o *corpus* é etiquetado e após são extraídas as palavras que combinavam com as características apresentadas (N e N N).

Geralmente, nas listas de unigramas e bigramas resultantes da aplicação do método híbrido, muitos termos possuem baixa freqüência ou muitos termos que aparecem em uma determinada lista de unigramas também fazem parte de algum bigrama da lista de bigramas. Assim, para encontrar unidades terminológicas mais apropriadas do domínio, são propostas algumas heurísticas para reduzir o número de termos identificados. Nessas heurísticas é utilizado o parâmetro Alpha, o qual, a partir da lista de unigramas e da lista de bigramas, permite favorecer a escolha de um unigrama ou de um bigrama para fazer parte da lista. Depois de aplicadas as heurísticas, são removidos da lista de termos candidatos todos os termos (unigramas ou bigramas) que possuem freqüência menor ou igual a um limiar Theta, definido pelo usuário, em relação ao número de documentos. A lista é utilizada na próxima fase do método.

Identificação de Atributos, Construção do Dicionário e da Tabela – A identificação dos atributos é realizada em três etapas: definição dos termos que serão utilizados como raiz das árvores[3]; geração das árvores; e identificação dos atributos a partir das árvores geradas. A identificação de termos raiz pode ser realizada de maneira automática ou não automática. No modo automático, todos os termos da lista de termos candidatos identificados na fase de ET são considerados para serem utilizados como termos raiz das árvores na próxima etapa. No modo não automático, uma análise, junto com os especialistas, pode ser realizada sobre a lista de termos candidatos final com o intuito de identificar os termos que realmente serão utilizados no mapeamento de informações, pois podem existir na lista termos que não são de interesse. Depois de definidos os termos raiz das árvores, é executado o algoritmo de geração de árvores, lembrando que para cada termo considerado como unidade terminológica é gerada uma árvore cuja raiz é definida por esse termo. As árvores geradas possuem uma estrutura semelhante à árvore ilustrada na Figura 2.

Nessa árvore, o termo α, onde o tamanho de α é dado por $1 \leq |\alpha| \leq 2$, corresponde ao nó raiz identificado pelo método de extração de terminologia, os termos β e γ são filhos de α e δ é filho de β. Na árvore, todos os filhos possuem um número de palavras maior ou igual a 1 e mapeiam as palavras que aparecem no contexto de um termo

[3]Em uma árvore são mapeados os termos que aparecem no contexto de uma unidade terminológica X, por exemplo, X A B, X C D, juntamente com a frequência desses termos. A partir das árvores são identificados os atributos que farão parte da tabela atributo-valor.

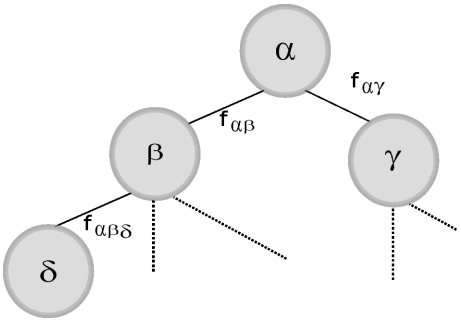

Figura 2. Árvore gerada

raiz. Por exemplo, se a palavra `coloração` e `esbranquiçada` sempre aparecem juntas, elas serão colocadas em um mesmo nó filho, caso contrário serão colocadas em nós diferentes. Na árvore gerada também são armazenadas as freqüências em que dois termos ocorrem juntos. Por exemplo, a freqüência com que α e β aparecem juntos é $f_{\alpha\beta}$, já a freqüência com que $\alpha\beta\delta$ aparecem juntos é $f_{\alpha\beta\delta}$, sendo $f_{\alpha\beta\delta} < f_{\alpha\beta}$. Essa relação entre as freqüências se verifica em todos os ramos da árvore. Na Figura 3 é ilustrada uma das interfaces do ambiente TP-DISCOVER, na qual podem ser visualizadas as árvores geradas a partir dos termos raiz.

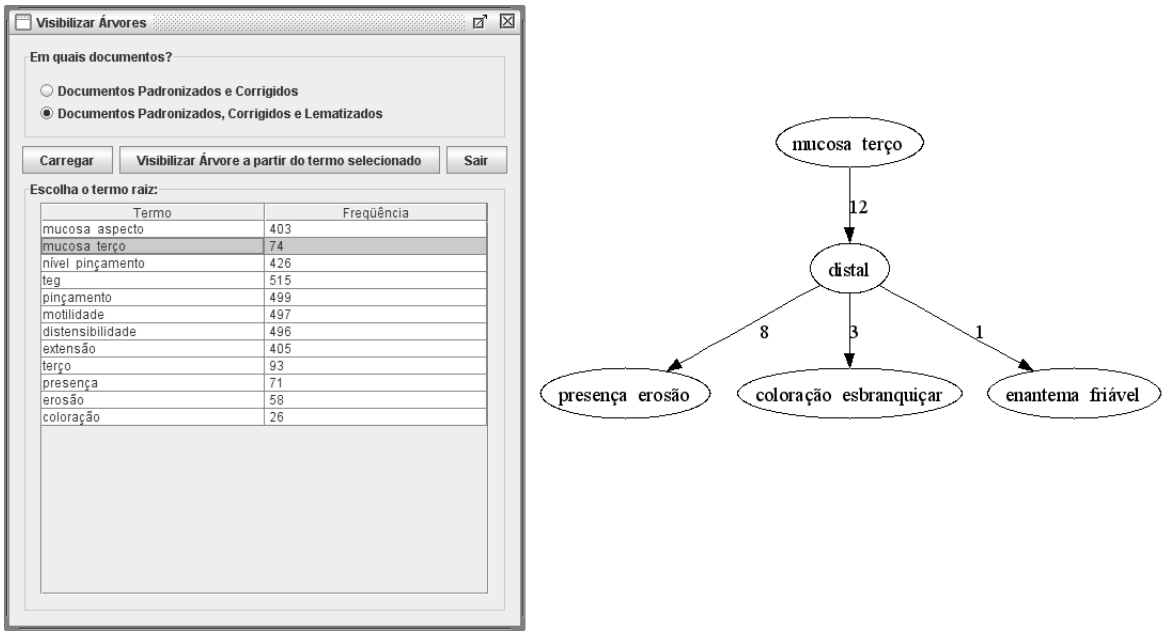

Figura 3. Interface para visualizar árvores

Após a construção do conjunto de árvores, a identificação dos atributos para compor a tabela atributo-valor pode ser realizada com o auxílio dos especialistas (modo não automático) ou automaticamente. No modo automático, todos os ramos que possuirem freqüência maior ou igual a um limiar são considerados atributos. Por exemplo, considere o limiar l e a árvore da Figura 2. Na identificação automática de atributos, primeiramente é verificado se $f_{\alpha\beta} \geq l$ e, se for, é gerado o atributo $\alpha\beta$. Depois é verificado se $f_{\alpha\beta\delta} \geq l$ e, se for, é definido $\alpha\beta\delta$ como atributo. Por último é verificado se $f_{\alpha\gamma} \geq l$ e, se for, é definido $\alpha\gamma$ como atributo. Depois de identificados os atributos, eles são

inseridos em um dicionário, denominado de dicionário de conhecimento. Ao final, o processo de preenchimento da tabela é realizado por meio de ciclos de pesquisa entre as informações dos documentos e os atributos presentes no dicionário de conhecimento construído. Cada documento corresponde a uma linha da tabela atributo-valor. Nesse processo, se a seqüência de termos definida por um determinado atributo do dicionário for identificada no documento, o valor desse atributo é preenchido com 1 (presente). Se a seqüência não for identificada no documento, o atributo é preenchido com 0 (ausente). Esse processo é repetido para todos os documentos do conjunto.

3. Descrição do Conjunto de Laudos

O objetivo deste trabalho é apresentar os resultados obtidos da aplicaçao do método a um conjunto de laudos de Endoscopia Digestiva Alta. Os laudos utilizados neste trabalho foram obtidos por meio de parcerias do — LABI — Laboratório de Bioinformática da Universidade Estadual do Oeste do Paraná com o Hospital Municipal de Paulínia em São Paulo. Todos os laudos estão em língua portuguesa, em formato digital armazenados em arquivos do tipo TXT e não possuem informações de referência, como nome do paciente ou médico que realizou o exame. Os laudos utilizados possuem o formato ilustrado na Figura 4.

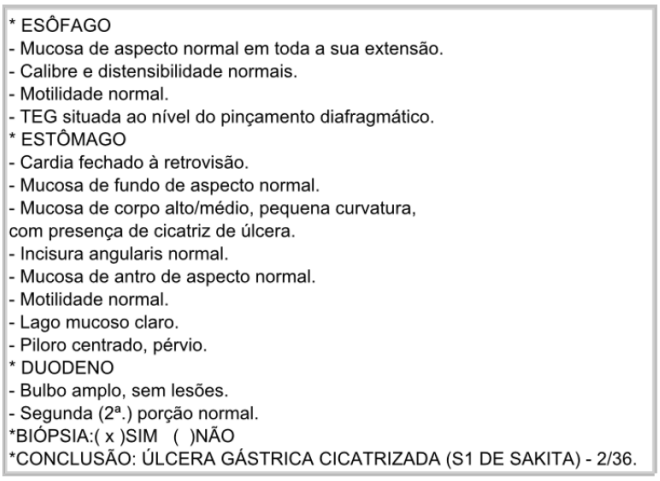
* ESÔFAGO
- Mucosa de aspecto normal em toda a sua extensão.
- Calibre e distensibilidade normais.
- Motilidade normal.
- TEG situada ao nível do pinçamento diafragmático.
* ESTÔMAGO
- Cardia fechado à retrovisão.
- Mucosa de fundo de aspecto normal.
- Mucosa de corpo alto/médio, pequena curvatura,
com presença de cicatriz de úlcera.
- Incisura angularis normal.
- Mucosa de antro de aspecto normal.
- Motilidade normal.
- Lago mucoso claro.
- Piloro centrado, pérvio.
* DUODENO
- Bulbo amplo, sem lesões.
- Segunda (2ª.) porção normal.
*BIÓPSIA:(x)SIM ( )NÃO
*CONCLUSÃO: ÚLCERA GÁSTRICA CICATRIZADA (S1 DE SAKITA) - 2/36.

Figura 4. Exemplo de laudo

Os laudos são organizados em cinco seções, contendo informações como propriedades e anormalidades do segmento de esôfago, estômago e duodeno, assim como informações referentes à realização de biópsia e conclusões do exame. Nas três primeiras seções e na última as informações estão descritas em língua natural. As informações sobre biópsia estão estrututras.

4. Avaliação Experimental

Foram realizados vários experimentos considerando o aspecto quantitativo, nos quais o método foi avaliado utilizando diferentes valores dos parâmetros. Também foi realizada uma avaliação de qualidade da lista de termos identificados pelo método híbrido de extração de terminologia desenvolvido. O método foi aplicado a um conjunto de laudos médicos composto por 6000 laudos de EDA. Conforme mencionado, neste trabalho utilizamos informações da seção de estômago dos laudos. Foi decidido avaliar os resultados

da aplicação do método sem a intervenção dos especialistas (modo automático), ou seja, considerando apenas o aspecto morfo-sintático. Assim, sem dúvida alguma, os resultados experimentais referem-se a resultados obtidos no *pior caso*, pois o aspecto semântico não foi considerado. Entretando, como os resultados necessitam ser avaliados não somente quantitativamente mas também qualitativamente, o conhecimento dos especialistas somente foi utilizado para calcular os valores de precisão e *recall* na fase de extração de terminologia, para os quais é necessário conhecer os termos de interesse do domínio.

Para avaliar o método proposto no pior caso, foram realizados vários experimentos utilizando quatro diferentes conjuntos de documentos relacionados ao estômago, os quais foram amostrados com reposição do conjunto total de 6000 documentos disponíveis. Na Tabela 1 é mostrado o número de documentos no conjunto de treinamento (Tr) e teste (Te) de cada um desses quatro experimentos. Na fase de extração de terminologia foram uti-

Tabela 1. Configuração dos experimentos realizados

Id. Experimento	Tr	Te
Exp1	500	1000
Exp2	1000	1000
Exp3	2000	1000
Exp4	4000	1000

lizados diferentes valores para os parâmetros Alpha e Theta a fim de observar a influência deles na identificação dos termos de interesse do conjunto de laudos. Para cada experimento identificado na Tabela 1, foram utilizadas nove variações dos valores de Alpha e Theta. Para o parâmetro Alpha foram utilizadas três variações, 100%, 95% e 90%. Para cada Alpha foram utilizadas três variações de Theta, 5%, 10% e 20%, ou seja, no primeiro experimento foram utilizados os valores de Alpha=100% e Theta=5%. No segundo foram utilizados Alpha=100% e Theta=10% e assim sucessivamente. Os experimentos foram realizados de modo a extrair as seguintes informações: número de termos raiz; número de atributos gerados utilizando limiares de poda da árvore iguais a 0%, 5%, 10% e 20%; e taxa de preenchimento da tabela atributo-valor.

A primeira tarefa antes de gerar as árvores foi realizar a identificação de unidades terminológicas na fase de extração de terminologia. A partir do conjunto de documentos etiquetados foram extraídos os termos pertencentes à classe gramatical N e N N. Sobre a lista de termos extraídos foram aplicadas as heurísticas que usam os valores de Alpha e Theta correspondentes. Após realizar a identificação desses termos (unidades terminológicas), os mesmos foram utilizados como termos raiz para gerar as árvores a partir das quais são identificados os atributos. Depois de geradas as árvores, foi realizada a identificação de atributos que compõem a tabela atributo-valor. Nos experimentos foram utilizados os limiares de poda das árvores 0%, 5%, 10% e 20%. Observar que utilizar o limiar 0% corresponde a transformar todos os ramos das árvores geradas em nomes de atributos. Na Figura 5 é ilustrado o gráfico que mostra a relação entre o número de termos raiz identificados e o número de atributos gerados utilizando o limiar de poda 0%, para os experimentos realizados.

Nesse gráfico, no eixo x estão representados os resultados obtidos com os valores de Alpha 100% (*a*), 95% (*b*) e 90% (*c*) e, para cada valor de Alpha estão representadas as variações de Theta (5%, 10%, 20%). É possível observar que usando as variações *a*, *b* e *c* de Alpha e Theta os resultados não variaram muito. Isso indica que as listas de termos raiz identificadas não foram muito diferentes. É possível também observar no

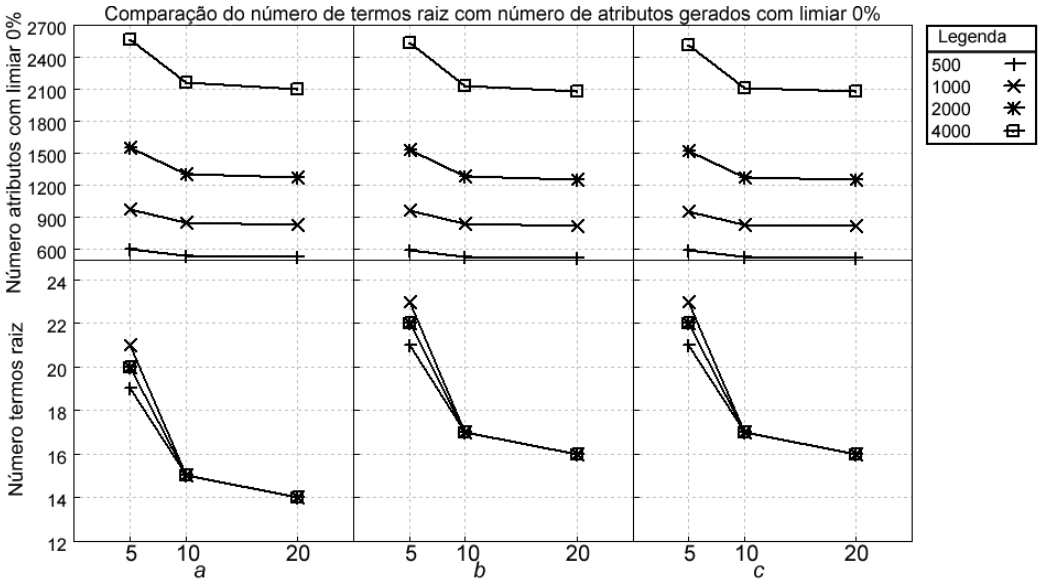

Figura 5. Relação entre número de termos raiz e número de atributos gerados utilizando limiar igual a 0% [Honorato 2008]

gráfico que quanto maior o número de documentos do conjunto de treinamento, maior o número de atributos identificados. Essa característica deve-se ao fato de que quanto mais documentos, existirão mais variações no contexto de um determinado termo raiz e, consequentemente, mais ramos serão criados a partir desse termo.

Para avaliar o número de atributos que são considerados em relação ao número total identificado com o limiar de poda igual a 0%, foram utilizados os limiares de 5%, 10% e 20%[4]. Nas Figuras 6 e 7 são ilustrados os gráficos que mostram o número de atributos considerados usando esses limiares.

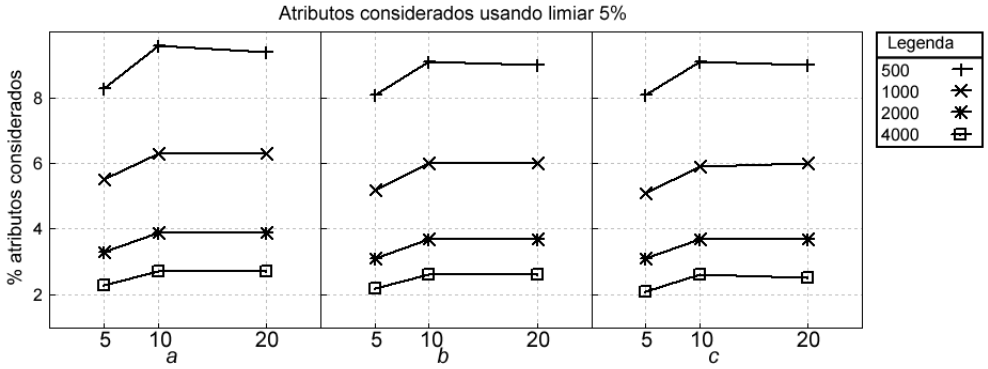

Figura 6. Taxa de atributos considerados utilizando limiar igual a 5% [Honorato 2008]

Utilizado o limiar de 5%, a taxa de atributos considerados em relação ao número total gerado é maior, uma vez que grande parte dos ramos das árvores possuem freqüência maior ou igual que 5%. É importante observar que para o conjunto de treinamento de 4000 laudos, a taxa de atributos considerados é menor pois, conforme observado na Figura 5, para esse experimento foi gerado um número muito maior de atributos, o que indica que

[4]Devido a questão de espaço são apresentados somente os gráficos dos limiares 5% e 20%.

foram geradas árvores com muitos ramos de freqüência menor que 5%. Por outro lado, a taxa de atributos considerados para o conjunto de treinamento de 500 laudos é maior, uma vez que o número de atributos gerado usando o limiar 0% é menor.

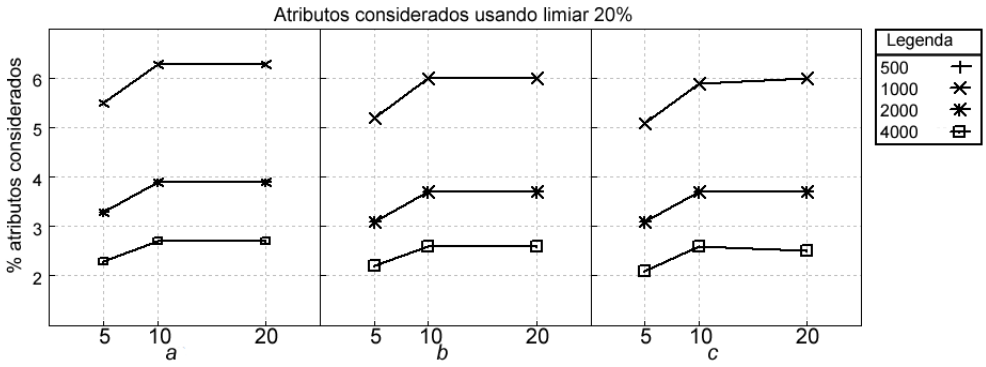

Figura 7. Taxa de atributos considerados utilizando limiar igual a 20% [Honorato 2008]

Embora a taxa de atributos considerados de um experimento para outro seja grande, a variação do número de atributos não é grande. Conforme mencionado, isso se deve ao fato de que, com um conjunto maior de documentos para treinamento, são geradas árvores com muitos ramos com freqüência baixa, enquanto os ramos com freqüência maiores estão próximos à raiz da árvore. Por outro lado, no caso na qual são considerados somente ramos que possuem freqüência maior ou igual a 10%, o número de atributos considerados é menor, uma vez que estão sendo considerados os ramos com freqüência maior ou igual que 10% do número de laudos. Isso se confirma no gráfico da Figura 7 (limiar de poda de 20%), no qual o número de atributos considerados é menor que para os limiares de 5% e 10%.

Após identificados os atributos utilizando os diferentes limiares (0%, 5%, 10% e 20%) foi realizada a construção do dicionário e o preenchimento da tabela atributo-valor. Usando limiar 0%, a taxa de preenchimento foi bastante baixa, ou seja, foi gerada uma tabela bastante esparsa. Por outro lado, conforme o valor do limiar de poda aumenta (5%, 10% e 20%), a taxa de preenchimento é incrementada, uma vez que o número de atributos é menor quando se usa limiares maiores, ou seja, estão sendo considerados como atributos apenas seqüências de palavras que têm freqüências maiores.

Quanto ao aspecto qualitativo, foram realizados os cálculos de precisão e *recall* das listas de termos geradas com base em uma lista de referência considerando apenas as informações contidas nos documentos da seção do estômago, os quais foram fornecidos pelos especialistas do domínio. Novamente foi observado que não ocorreram variações muito grandes nos valores de precisão e *recall* entre os diferentes experimentos. No entanto, foi alcançada melhor precisão (69%, com Theta=20%) e *recall* (57%, com Theta=5%) usando Alpha = 95% e Alpha = 90%, em todos os experimentos. Analisando a lista de referência, notou-se que muitos termos de interesse utilizados na seção do estômago são bigramas e, usando Alpha = 95% e Alpha = 90% a identificação desse tipo de termo é favorecida. Deve ser levado em conta que devido a utilização de um vocabulário controlado, o uso de conjuntos de laudos de tamanhos diferentes para a extração de terminologia de interesse não influenciou muito em relação a melhores resultados de

precisão e *recall*.

5. Considerações Finais

O objetivo deste trabalho foi apresentar os resultados experimentais da aplicação do método proposto por [Honorato 2008] no modo automático. O método foi avaliado no *pior caso*, ou seja, considerando apenas o aspecto morfo-sintático das informações, mostrando bons resultados. Se os especialistas participarem do processo (modo não automático), o aspecto semântico pode ser considerado e os resultados serão certamente superiores. Quanto aos resultados qualitativos, a precisão e o *recall* obtidos nos diversos experimentos realizados mostram a adequabilidade do método utilizado. Outro aspecto importante a ser ressaltado é que para aplicar o método não é necessário utilizar recursos externos de conhecimento do domínio tais como dicionários de termos, regras semânticas e ontologias do domínio. Somente é necessário conhecimento externo, dos especialistas, caso for utilizado o modo não automático. Em outras palavras, o método utiliza somente informações contidas nos documentos.

Agradecimentos: Trabalho realizado com o auxílio do Programa de Desenvolvimento Tecnológico Avançado — PDTA da Fundação Parque Tecnológico Itaipu — FPTI-BR e do Conselho Nacional de Desenvolvimento Científico e Tecnológico — CNPq.

Referências

Feldman, R. and Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, Nova Iorque, EUA.

Honorato, D. D. F. (2008). Metodologia de transformação de laudos médicos não estruturados e estruturados em uma representação atributo-valor. Dissertação de Mestrado, ICMC-USP, http://www.teses.usp.br/teses/disponiveis/55/55134/tde-10062008-154826. Acesso em 18/06/2009.

Honorato, D. D. F., Cherman, E. A., Lee, H. D., Monard, M. C., and Wu, F. C. (2008a). Construction of an attribute-value representation for semi-structured medical findings knowledge extraction. *CLEI Electronic Journal*, 11(2):1–12.

Honorato, D. D. F. and Monard, M. C. (2008). Descrição do ambiente computacional TP-DISCOVER para mapear informações não estruturadas em uma tabela atributo-valor. Technical Report 318, ICMC-USP, http://www.icmc.usp.br/b̃iblio/BIBLIOTECA/rel_tec/RT_318.pdf. Acesso em 18/06/2009.

Honorato, D. D. F., Monard, M. C., Lee, H. D., and Wu, F. C. (2008b). Uma abordagem de extração de terminologia para a construção de uma representação atributo-valor a partir de documentos não estruturados. In *Conferencia Latinoamericana de Informática*, pages 190–199, Santa Fe, Argentina.

Lee, H. D. (2005). Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado, ICMC-USP, http://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219. Acesso em 18/06/2009.

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufman Publishers Inc., San Francisco, Califórina, EUA.