Um Estudo do Mapeamento de Laudos Médicos de Endoscopia Digestiva Alta e Colonoscopia para Aquisição de Conhecimento

Everton Alvares Cherman[1], Newton Spolaôr[1], Huei Diana Lee[1,2], Daniel de Faveri Honorato[2], Cláudio Sadi Rodrigues Coy[3], João José Fagundes[3], Feng Chung Wu[1,2,3]

[1]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85856-970 – Foz do Iguaçu, PR, Brasil

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

[3]Faculdade de Ciências Médicas – Universidade Estadual de Campinas
Serviço de Coloproctologia
Caixa Postal 6111, 13083-970 – Campinas, SP, Brasil

{evertoncherman,newtonspolaor,hueidianalee}@gmail.com

Abstract. *The Data Mining process may help specialists on decision making by applying patterns extraction techniques based on attribute-value tables. An automatic medical report information mapping method is being developed intending to reduce the necessary time of the process and to avoid possible subjectivity on the manual information mapping. This work presents a broad case study about this method using information from 100 colonoscopy medical reports and on 609 upper digestive endoscopy medical reports, from which 82% and 100% were, respectively, automatically mapped.*

Resumo. *O processo de Mineração de Dados pode auxiliar especialistas no processo de tomada de decisão, por meio da extração de padrões a partir de tabelas atributo-valor. Uma metodologia de mapeamento automático de informações de laudos médicos está sendo desenvolvida com o intuito de reduzir o tempo necessário para o processo e evitar uma possível subjetividade do mapeamento manual dessas informações. Este trabalho apresenta um estudo de caso amplo sobre a aplicação desta metodologia sobre informações presentes em 100 laudos médicos colonoscópicos e em 609 laudos médicos de endoscopia digestiva alta, dos quais 82% e 100% foram, respectivamente, mapeados automaticamente.*

1. Introdução

O crescimento contínuo da utilização de tecnologia para aquisição e armazenamento de dados tem permitido o acúmulo de dados em uma velocidade maior que a capacidade humana possui para processá-los. Na área médica, assim como em outras áreas, esse crescimento é perceptível, uma vez que uma quantidade considerável das informações de

pacientes estão descritas em laudos e formulários médicos no formato eletrônico. 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 é inviável, pois trata-se de uma tarefa que tem alto custo de tempo e que está sujeita à subjetividade [1, 2, 3, 4]. Desse modo, para que esses dados textuais brutos possam tornar-se úteis, é necessário que eles sejam representados de maneira apropriada. Então esses dados poderão ser processados para extrair padrões, tal que um modelo que represente o conhecimento embutido nesses dados seja construído. Uma das maneiras de alcançar esse objetivo é por meio da realização do processo de Mineração de Dados – MD [5], o qual é constituído, usualmente, pelas etapas de pré-processamento, extração de padrões e pós-processamento.

O pré-processamento tem como objetivo preparar, reduzir e transformar os dados para um formato adequado para a extração de padrões. Um dos formatos mais utilizados é o atributo-valor, no qual as linhas representam os casos (exemplos) e as colunas os valores de cada característica considerada (atributo). É importante ressaltar que essa é a etapa mais demorada, a qual consome em torno de 80% do tempo necessário para realizar o processo, pois deve assegurar que os dados sejam representativos para as próximas etapas [6].

A etapa de extração de padrões tem por objetivo construir modelos a partir da tabela atributo-valor – TAV. Nessa etapa podem ser utilizados, por exemplo, algoritmos de inteligência artificial da área de aprendizado de máquina. Os modelos identificados podem ser representados por estruturas simbólicas como árvores de decisão e regras de produção, as quais permitem maior compreensibilidade humana [7].

Os padrões observados anteriormente são avaliados e validados com o auxílio de especialistas do domínio na etapa de pós-processamento. Os modelos consolidados possibilitam constituir novo conhecimento, o qual pode contribuir com o processo de tomada de decisão [8].

Conforme mencionado, tecnologias de armazenamento estão sendo cada vez mais utilizadas para registrar informações de pacientes. Essas informações, relacionadas a prognósticos e diagnósticos de exames nas diversas especialidades médicas são armazenadas, geralmente, em Laudos Médicos – LM – semi-estruturados descritos em língua natural. No contexto deste trabalho, as informações armazenadas nos laudos de Endoscopia Digestiva Alta – EDA – estão relacionadas às propriedades e anormalidades do esôfago, estômago e duodeno e nos laudos de colonoscopia são armazenadas informações relacionadas à descrição das condições patológicas do intestino grosso.

Para que possa ser aplicado o processo de MD sobre as informações armazenadas nos LM apresentados, é necessário que essas informações sejam transformadas para o formato adequado utilizado por esses algoritmos, geralmente o formato atributo-valor. A área de extração de informação [9] pode auxiliar nessa tarefa, por meio de métodos que, baseados em restrições sintáticas e semânticas, realizam a construção de representações estruturadas a partir de textos não estruturados em língua natural que possuem uma gramática bem definida. Na literatura podem ser encontrados alguns trabalhos da área de extração de informação [10, 11, 12, 13], os quais utilizam diferentes técnicas para transformação de informações não estruturadas contidas em registros médicos em rep-

resentação estruturada. Neste trabalho consideramos laudos médicos semi-estruturados mas que não possuem uma gramática bem definida. Mais especificamente, é apresentado um estudo amplo neste trabalho, utilizando conjuntos de LM de EDA e de colonoscopia, da metodologia proposta em [2, 4], a qual tem por objetivo dar suporte à construção de uma tabela atributo-valor a partir de LM semi-estruturados descritos em língua natural. Alguns estudos foram realizados com sucesso utilizando essa metodologia [14, 4, 15, 16].

Este trabalho está inserido no projeto de Análise Inteligente de Dados, o qual é desenvolvido por meio de uma parceria entre o Laboratório de Bioinformática – LABI – da Universidade Estadual do Oeste do Paraná – UNIOESTE/Foz do Iguaçu – , o Laboratório de Inteligência Computacional – LABIC – da Universidade de São Paulo – USP/São Carlos – e o Serviço de Coloproctologia da Universidade Estadual de Campinas – UNICAMP.

O restante deste trabalho está organizado da seguinte maneira: na Seção 2 é descrita a metodologia proposta bem como os LM considerados; na Seção 3 são apresentados e discutidos os resultados do trabalho e na Seção 4 são descritas as conclusões e os trabalhos futuros.

2. Materiais e Métodos

É consensual que doenças do sistema digestivo, como úlceras e câncer colorretal, apresentam alta incidência na população mundial. Nesse sentido, os exames de EDA e de colonoscopia contribuem no diagnóstico de enfermidades esofagogastroduodenais e colorretais, respectivamente [17, 18, 19].

Neste trabalho foram considerados 609 LM de EDA, confeccionados pelo Serviço de Endoscopia Digestiva do Hospital Municipal de Paulínia, que apresentam as informações organizadas em quatro segmentos, nos quais os três primeiros apresentam informações do esôfago, do estômago e do duodeno. O último segmento apresenta observações importantes, como a conclusão formulada pelo médico, e resultados de outros exames complementares, como patologia e teste da urease.

Os 100 LM de colonoscopia, considerados neste trabalho e confeccionados pelo Serviço de Coloproctologia da UNICAMP, são compostos por um segmento com campos previamente definidos e um com texto escrito livremente. O segmento estruturado contém dados do paciente, detalhes técnicos do exame e outras observações. Na porção desestruturada desses LM estão descritas informações correspondentes ao exame de colonoscopia.

A metodologia proposta em [2, 4] e aplicada nesses LM é constituída por duas fases, as quais são representadas pela Figura 1 e descritas a seguir.

A primeira fase é realizada por meio das etapas de “Identificação de Padrões” e de “Construção do Dicionário” e tem como objetivo final a construção de um dicionário do domínio, o qual auxilia no processo de mapeamento dos LM. Na primeira etapa, os padrões textuais presentes nos LM são detectados por meio de quatro tarefas:

1. Identificação de frases únicas.
2. Definição da lista de *stopwords*.
3. Construção de um Arquivo de Padronização – AP.
4. Geração de *n*-gramas.

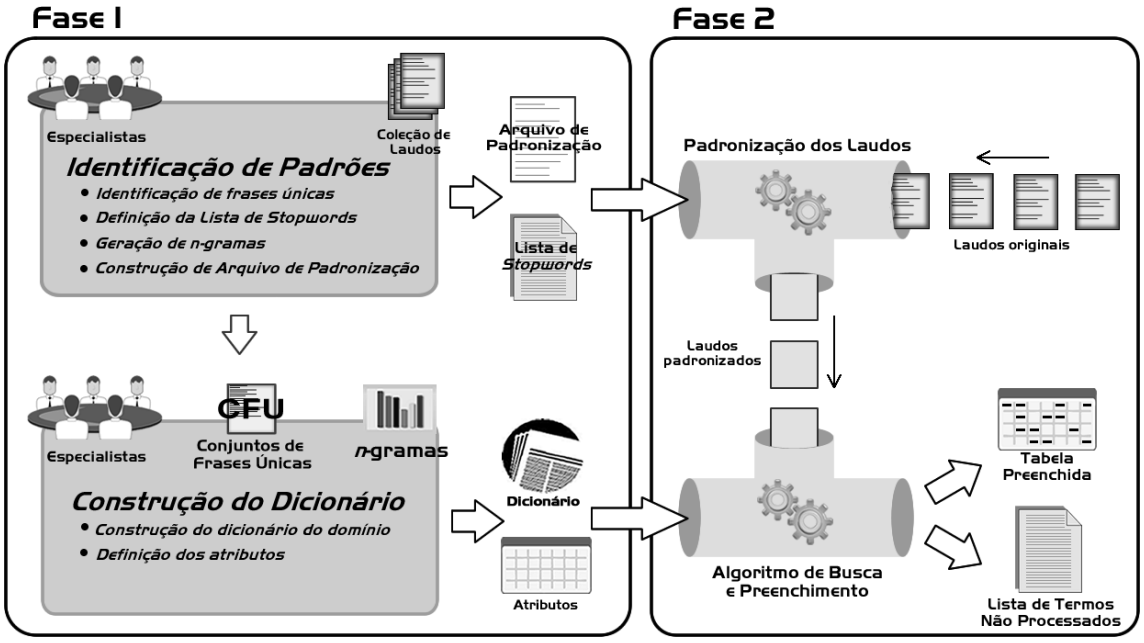

Figura 1. Representação das duas fases da metodologia

Na tarefa 1 são constituídos quatro conjuntos de frases únicas. Cada conjunto é construído agrupando todas as frases contidas no conjunto de LM em apenas um documento e retirando as frases repetidas. A diferença entre os quatro conjuntos é o nível de padronização embutido nesses conjuntos. Segue abaixo a descrição de cada conjunto:

1. O conjunto que contém as frases únicas originais dos LM, isto é, sem padronizações, é considerado o primeiro Conjunto de Frases Únicas – CFU1.
2. O segundo Conjunto de Frases Únicas – CFU2 – é definido a partir da Remoção de *Stopwords* – RS – sobre o CFU1. Ao realizar-se esse procedimento, retira-se das frases as preposições, os artigos, as conjunções e também algumas palavras definidas conjuntamente com especialistas. Essa palavras são as que ao serem eliminadas não modificam o sentido original das frases. Desse modo, é possível construir um conjunto de frases únicas com menor redundância, o que diminuí a quantidade de frases a serem analisadas.
3. O terceiro Conjunto de Frases Únicas – CFU3 – é construído a partir da Aplicação de *Stemming* – AS – sobre o CFU2. A técnica de AS tem como objetivo substituir por um radical comum as palavras que se diferenciam apenas por suas diferentes inflexões. Com isso é possível retirar possíveis redundâncias que não foram identificadas no CFU2.
4. Opcionalmente, é possível aplicar a técnica de lematização sobre o CFU2 e construir um quarto Conjunto de Frases Únicas – CFU4. A lematização tem o mesmo objetivo da AS: eliminar diferentes inflexões. No entanto, as palavras resultantes dessa técnica constituem frases com maior legibilidade em relação ao CFU3, pois as palavras são transformadas para sua forma canônica, como o infinitivo de um verbo e o masculino e singular de um substantivo.

A tarefa 2 na primeira etapa tem como objetivo definir uma lista de *Stopwords*. A construção dessa lista é iniciada ainda na primeira tarefa para a elaboração do CFU2 e é atualizada conjuntamente com os especialistas continuamente até o fim da primeira etapa.

A freqüente utilização de sinônimos na descrição de informações semelhantes presentes nos laudos médicos ou a presença de frases que expressam informações de uma maneira diferente da que será utilizada pelo dicionário, faz com que a padronização das informações contidas nos LM seja necessária. A construção do AP pode ser iniciada em paralelo à tarefa 1 em conjunto com especialistas e continua até o fim da primeira etapa. A aplicação da padronização permitirá que as informações contidas nos laudos estejam mapeadas em um formato padrão para serem utilizadas pelo dicionário e pelo processo de preenchimento da base de dados. Na Figura 2 são apresentados dois exemplos de padronizações.

Antes da padronização		Após a padronização
coloração esbranquiçada	➡	anormal
calibre e distensibilidade normais	➡	calibre normal distensibilidade normal

Figura 2. Exemplos de padronizações de frases dos LM

Na tarefa 4, com o intuito de auxiliar na identificação de padrões, é realizada a geração de *n*-gramas sobre os LM. Um *n*-grama é definido como a freqüência em que *n* palavras consecutivas estão descritas no conjunto de documentos. Uma lista de 1-grama é constituída por todas as palavras presentes nos LM seguidas da freqüência em que são descritas. Assim como uma lista de 2-gramas contém a freqüência com que todas as combinações de duas palavras consecutivas são descritas nos documentos. Desse modo, é possível identificar unidades terminológicas utilizadas no domínio de interesse, os quais possivelmente contém uma freqüência maior.

A segunda etapa, construção do dicionário, é realizada em conjunto com os especialistas e com o auxílio dos artefatos gerados na primeira etapa, para a construção do dicionário e a definição dos atributos que integrarão a Tabela Atributo-Valor – TAV.

Na maioria das especialidades médicas são descritas nos LM informações sobre as estruturas anatômicas examinadas e suas respectivas características. Por exemplo, nos laudos de EDA, na seqüência “terço distal com erosões” o termo “terço distal” é a estrutura anatômica que está em análise e o termo “com erosões” é a característica associada com esse local. Em ambos exames tratados neste trabalho são registradas opcionalmente subcaracterísticas, ou seja, particularidades de uma característica.

Desse modo, o dicionário é formado por uma estrutura hierárquica de três níveis, composta por locais, características e subcaracterísticas, os quais correspondem, respectivamente, às estruturas anatômicas, características e particularidades dessas características presentes nos LM. Para cada local representado no dicionário, existe uma lista de *n* características e cada uma, por sua vez, exibe uma lista de *m* subcaracterísticas. Os atributos e os valores da TAV são formados com base nas relações entre locais e características ou entre locais, características e subcaracterísticas identificadas. Por exemplo, o texto “terço distal com erosões” forma uma relação de local (terço distal) e característica (com erosões). Nesse caso, seria criado um atributo com o nome de “terço distal” e um dos valores seria “com erosão”.

O mapeamento das informações dos LM para a TAV é o foco da segunda fase da metodologia. Na construção da TAV é utilizado o conjunto de laudos, o AP e o dicionário. O processo de mapeamento é realizado pelo Algoritmo de Busca e Preenchimento – ABP. Esse algoritmo é executado da seguinte maneira: a partir de um conjunto de LM inicial, um exemplar é extraído para padronização baseada no AP; posteriormente, são consultados em cada uma das frases desse exemplar todos os locais definidos no dicionário; quando um local está presente, as características relacionadas a este local são também consultadas e se alguma característica for encontrada, é realizada uma consulta pelas subcaracterísticas correspondentes a essa característica.

Conforme mencionado, as relações entre locais e características, ou entre locais, características e subcaracterísticas, presentes em um LM definem os atributos e seus respectivos valores, os quais constituem um exemplo da TAV. Essa instância, referente ao exemplar sob mapeamento, é armazenada na TAV, e o processo se reinicia a partir de outro exemplar extraído do conjunto inicial de LM.

3. Resultados e Discussão

Conforme mencionado, neste trabalho a metodologia foi aplicada a laudos de dois domínios diferentes, EDA e colonoscopia. Os laudos médicos de EDA foram segmentados em três partes, esôfago, estômago e duodeno e, a metodologia foi aplicada a cada um desses segmentos separadamente.

3.1. Primeira Fase

Os resultados obtidos a partir da aplicação da metodologia descrita anteriormente na primeira fase são apresentados na Tabela 1.

Tabela 1. Resultados da primeira fase

Domínio	EDA – Esôfago	EDA – Estômago	EDA – Duodeno	Colonoscopia
Total frases	3044	5226	1710	474
CFU1	81	352	90	412
CFU2	65	259	88	396
CFU3	64	257	81	393
Padronizações	56	43	30	274
Atributos	16	22	51	277

Nessa tabela, as colunas correspondem aos conjuntos de laudos aos quais a metodologia foi aplicada e as quatro primeiras linhas correspondem aos números de frases totais de cada conjunto de laudos e número de frases do CFU1, CFU2, CFU3, respectivamente. A quinta linha corresponde ao número de padronizações identificadas e a última linha corresponde ao número de atributos que foram identificados para cada conjunto de laudos.

Na primeira fase aplicada à porção do esôfago foi possível identificar inicialmente 3044 frases. Desse total, apenas 81 frases únicas constituíram o CFU1, o que representa uma redução de 97,34% em relação ao número total de frases. A utilização da RS em CFU1 possibilitou a geração do CFU2, o qual foi composto por 65 frases únicas, reduzindo em 19,75% a quantidade de frases contida em CFU1. Em seguida, ao aplicar a técnica de *stemming* no CFU2, foi gerado o CFU3 contendo 64 frases, o que permitiu

reduzir em 20,99% a quantidade de frases de CFU1. O AP foi composto por 56 padronizações e a TAV foi definida contendo 16 atributos.

A porção do estômago continha um total de 5226 frases, das quais 352 compuseram o CFU1, representando uma redução de 93,26%. O CFU2 foi composto por 259 frases e o CFU3 por 257 frases, o que representa, respectivamente, reduções de 26,42% e de 26,99% em relação ao CFU1. O AP foi constituído por 43 registros e a TAV foi definida com 22 atributos.

As informações referentes à porção do duodeno estavam representadas em 1710 frases. O número de frases foi reduzido em 94,74% no CFU1, o qual foi composto por 90 frases. O CFU2 foi composto por 88 frases e o CFU3 por 81 frases, o que representa, respectivamente, reduções de 2,22% e de 10% em relação ao CFU1. No AP foram definidas 30 padronizações e a TAV desse domínio foi composta por 51 atributos.

No domínio da colonoscopia foi possível coletar do conjunto de 100 LM um total de 474 frases. Dessas frases, 412 compuseram o CFU1, reduzindo em 13,08% a quantidade de frases. A RS e a AS sobre o CFU1 resultaram no CFU2 e no CFU3 contendo 396 e 393 frases respectivamente, apresentando reduções de 16,45% e 17,08% em relação ao total. Nesse domínio foram aplicadas as técnicas de lematização e de geração de *n*-gramas. A lematização aplicada ao CFU2 possibilitou gerar o CFU4 com a mesma quantidade de frases do CFU3, porém com frases de maior legibilidade. A geração de n-gramas sobre os laudos resultou em 1812 bigramas e 2321 trigramas, dos quais 129 e 52 respectivamente foram identificados em cinco ou mais LM, o que possibilitou confirmar alguns termos do domínio pertencentes ao dicionário. O AP foi composto por 274 padronizações e a TAV definida com 277 atributos.

Uma avaliação dos resultados, mostrou que no domínio do EDA, os CFU1 foram gerados com um número significantemente menor em relação ao número total de frases. Isso indica que há uniformidade na descrição das informações e conforme foi constatado, os laudos foram gerados por um único médico. Por outro lado, pode-se notar que nos laudos de colonoscopia na construção do CFU1 não ocorreu uma redução significativa como nos laudos de EDA. Em uma análise mais detalhada foi constatado que os laudos de colonoscopia foram gerados por um número elevado de pessoas diferentes (médicos residentes) e, portanto, a uniformidade na descrição das informações foi menor.

3.2. Segunda Fase

Finalizada a construção do dicionário, foi iniciada a segunda fase. Nessa fase, os LM de EDA e de colonoscopia foram submetidos ao ABP para que as informações contidas nesses LM fossem mapeadas para as TAV correspondentes. Os termos não processados pelo ABP são registrados em um conjunto, denominado de conjunto de termos não processados, para auxiliar na avaliação da precisão do mapeamento automático realizado nessa fase. Utilizando esse conjunto, após o processamento dos conjuntos de LM, foi possível calcular a quantidade de valores efetivamente mapeados automaticamente através do ABP.

Na Tabela 2 são apresentados os resultados da aplicação da segunda fase da metodologia.

No domínio do EDA, para os 609 LM, foram apurados que 8760, 3340 e 9533 va-

Tabela 2. Valores de precisão do processamento para os quatro conjuntos de LM processados

Domínio	EDA – Esôfago	EDA – Estômago	EDA – Duodeno	Colonoscopia
Esperado	8760 valores	3340 valores	9533 valores	967 valores
Efetivo Automático	100%	100%	100%	82%
Efetivo Manual	0%	0%	0%	18%
Efetivo Total	100%	100%	100%	100%

lores de atributos foram efetivamente preenchidos para a porção do esôfago, do estômago e do duodeno, respectivamente. No conjunto de 100 LM de colonoscopia foram mapeados automaticamente 793 valores para a tabela atributo-valor.

No domínio da colonoscopia, a partir do conjunto de termos não processados, foi possível constatar, que 174 valores de atributos não foram mapeados para a TAV, ou seja, dos 967 valores esperados[1] para o mapeamento, 82% foram mapeados automaticamente. No domínio do EDA, 100% dos valores foram preenchidos para a TAV, isto é, não houve registros de palavras não processadas.

Em uma análise dos preenchimentos de ambos os domínios, observou-se que o principal fator que influenciou nesses resultados é a forma de descrição das informações nos LM, com a qual está fortemente relacionada o ABP. Conforme mencionado, no domínio do EDA, os LM foram preenchidos por um mesmo especialista, contendo uma descrição mais uniforme, o que diminui a complexidade necessária para o ABP. Essa característica de uniformidade estava menos presente no conjunto de LM de colonoscopia, os quais eram descritos por diversos especialistas. Esse fato demandou uma quantidade maior de padronizações em relação ao domínio de EDA, chegando a quase dez vezes mais padronizações que a porção do duodeno. Desse modo, também seria necessário um ABP mais complexo para se mapear integralmente as informações. Para que a tabela fosse integralmente preenchida, os valores que não foram mapeados automaticamente, foram preenchidos manualmente com auxílio do conjunto de termos não processados, completando 100% do mapeamento esperado para a TAV.

4. Conclusão

Neste trabalho é apresentado um amplo estudo de caso de mapeamento de LM para tabelas no formato atributo-valor utilizando uma metodologia de mapeamento automático, o qual envolveu laudos médicos dos domínios de colonoscopia e endoscopia digestiva alta.

Com a construção dos dicionários para cada domínio foi possível mapear os LM para a TAV de maneiras semi-automática, no caso da colonoscopia, e integralmente automática, no caso do EDA. Observou-se também que a uniformidade da descrição dos LM influenciou diretamente na precisão do mapeamento automático das informações, pois essa característica está fortemente relacionada ao ABP, o qual realiza o processamento dos LM. Contudo, após o mapeamento manual das informações não processadas, foi possível obter tabelas integralmente preenchidas com informações referentes aos exames de EDA e de colonoscopia, as quais serão disponibilizadas para a etapa de extração de padrões.

[1]Considerou-se valores esperados os valores de atributos os quais estão mapeados no dicionário construído.

Em uma avaliação conjunta com especialistas do domínio, foi observado que a metodologia auxiliou na redução, em relação ao mapeamento completamente manual, do tempo demandado na etapa de pré-processamento e possibilitou um preenchimento padronizado e não subjetivo das informações para a tabela. Outro aspecto importante a ser ressaltado é que com os dicionários construídos, será possível mapear novos conjuntos de LM das áreas de colonoscopia e EDA sem a necessidade de construir novamente esses dicionários. Se necesário esses dicionários poderão ser facilmente atualizados com a adição de novos atributos com o auxílio do conjunto de termos não processados.

Como trabalhos futuros pretende-se adequar a complexidade do ABP para LM de colonoscopia, a fim de atingir 100% de precisão no preenchimento automático, e também aplicar a etapa de extração de padrões às tabelas construídas. A estrutura de representação do conhecimento utilizado pelo dicionário é adequada para a interação com o ABP, no entanto, utilizar padrões amplamente considerados para representação, como os padrões de ontologias, pode prover ganhos em poder de representação e em interoperabilidade semântica. Nesse sentido, outro trabalho futuro corresponde ao estudo da viabilidade em adequar a atual estrutura do dicionário para padrões propostos na literatura e amplamente utilizados.

Agradecimentos

Ao Programa de Desenvolvimento Tecnológico Avançado – PDTA/FPTI-BR – e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq – pelo auxílio por meio da linha de financiamento de bolsas.

Referências

[1] H. D. Lee and M. C. Monard. Seleção de atributos para algoritmos de aprendizado de máquina
supervisionado utilizando como filtro a dimensão fractal. *Revista de La Sociedad Chilena de Ciencia de La Computación*, pages 1–8, 2003.

[2] D. D. F. Honorato, H. D. Lee, M. C. Monard, F. C. Wu, R. B. Machado, A. P. Neto, and C. A. Ferrero. Uma metodologia para auxiliar no processo de construção de bases de dados estruturadas a partir de laudos médicos. In *Anais do Encontro Nacional de Inteligência Artificial*, pages 593–601, São Leopoldo - RS, 2005.

[3] H. D. Lee. *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. Tese de Doutorado, Universidade de São Paulo, São Carlos, SP, 2005.

[4] D. D. F. Honorato, E. A. Cherman, H. D. Lee, M. C. Monard, and F. Wu. Construção de uma representação atributo-valor para extração de conhecimento a partir de informações semi-estruturadas de laudos médicos. In *Anais do Conferencia Latinoamericana de Informática*, San José - Costa Rica, 2007.

[5] U. M. Fayyad, G. Platestsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In *American Association for Artificial Intelligence*, pages 1–30, 1996.

[6] D. Pyle. *Data preparation for data mining*. The Morgan Kaufmann, 1999.

[7] I. H. Witten and E. Frank. *Data Mining Practical Machine Learning Tools and Techniques*. Elsevier, 2005.

[8] S. O. Rezende. *Sistemas Inteligentes - Fundamentos e Aplicações*. Manole, 2003.

[9] I. Muslea. Extraction patterns for information extraction tasks: A survey. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction*, Menlo Park, California, USA, 1999.

[10] D. Bekhouche, Y. Pollet, B. Grilheres, and X. Denis. Architecture of a medical information extraction system. In *Anais do 9th International Conference on Applications of Natural Language to Information Systems*, pages 380–387, Salford, UK, 2004.

[11] H. Harkema, I. Roberts, R. Gaizauskas, and M. Hepple. Information extraction from clinical records. In *Anais do the 4th UK e-Science All Hands Meeting*, Nottingham, UK, 2005.

[12] R. K. Taira, S. G. Soderland, and R. M. Fakobovits. Automatic structuring of radiology free-text reports. *Radiographics*, 21:237–245, 2001.

[13] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks. Approaches to text mining for clinical medical records. In *Anais do SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*, pages 235–239, New York, NY, USA, 2006. ACM Press.

[14] C. A. Ferrero, H. D. Lee, F. C. Wu, R. B. Machado, D. D. F. Honorato, J. J. Fagundes, and J. R. N. Góes. Um estudo de caso de construção de base de dados a partir de laudos médicos. In *Anais do 13º Simpósio Internacional de Iniciação Científica da USP* , São Carlos - SP, 2005.

[15] E. A. Cherman, H. D. Lee, D. D. F. Honorato, J. J. Fagundes, J. R. N. Góes, C. S. R. Coy, and F. C. Wu. Metodologia de mapeamento de laudos médicos para bases de dados: Aplicação em laudos colonoscópicos. In *Anais do II Congresso da Academia Trinacional de Ciências*, pages 1–9, Foz do Iguaçu - PR, 2007.

[16] N. Spolaôr, H. D. Lee, E. A. Cherman, D. D. F. Honorato, J. J. Fagundes, J. R. N. Góes, C. S. R. Coy, and F. C. Wu. Um estudo de caso do mapeamento de laudos endoscópicos para bases de dados. In *Anais do II Congresso da Academia Trinacional de Ciências*, pages 1–10, Foz do Iguaçu - PR, 2007.

[17] F. A. Quilici. *Colonoscopia*. Lemos Editorial, 2000.

[18] F. Cordeiro. *Endoscopia Digestiva*. Editora Média e Científica Ltda., 1994.

[19] R. S. Cotran, V. Kumar, and T. Collins. *Patologia Estrutural e Funcional*. Guanabara Koogan, 2000.