Construção de uma Ontologia para Auxiliar no Mapeamento de Laudos Médicos de Endoscopia Digestiva Alta para Bases de Dados Estruturadas

Luiz Henrique Dutra da Costa[1], Carlos Andrés Ferrero[1], Huei Diana Lee[1], Cláudio Saddy Rodrigues Coy[2], João José Fagundes[2], Feng Chung Wu[1,2]

[1]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85869-970 – Foz do Iguaçu, PR, Brasil

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

{lhdc90, anfer86, hueidianalee, wufengchung}@gmail.com

Resumo. *Na área médica, as informações são normalmente armazenadas em laudos textuais, os quais se encontram em formato semiestruturado ou não-estruturado. Porém, para que métodos computacionais de extração de conhecimento possam ser aplicados, os dados precisam estar num formato estruturado. Para essa tarefa, foi desenvolvido, em um trabalho prévio, um método de mapeamento de laudos médicos textuais para bases de dados estruturadas. Este trabalho tem como objetivo a construção de uma ontologia para estender esse método. A nova abordagem foi aplicada a um conjunto de laudos de Endoscopia Digestiva Alta permitindo mapear 100% dos atributos esperados e obter uma representação do conhecimento do domínio do laudo por meio da ontologia.*

1. Introdução

Devido à constante evolução tecnológica e à disseminação do uso de computadores em diversas áreas, um volume cada vez maior de informações é armazenado por meio de sistemas gerenciadores de dados. Desse modo, pode ser interessante analisar essas grandes bases de dados com o intuito de extrair algum conhecimento útil, porém devido ao seu tamanho crescente, a análise manual se torna inviável e de alto custo, sendo necessária a aplicação de diversos métodos computacionais para dar suporte a essa análise, assim como na construção de modelos que representam o conhecimento presente nesses dados. Um dos métodos que podem ser utilizados para dar apoio a essa tarefa é a Mineração de Dados — MD [Han and Kamber 2006].

O processo de MD pode ser dividido em três etapas [Rezende 2003]: pré-processamento, extração de padrões e pós-processamento. A etapa de pré-processamento tem como objetivo adequar os dados para a extração de padrões por meio de seleção, limpeza e preparação dos dados. A etapa de extração de padrões tem como objetivo construir modelos com base nos dados. E por último, a etapa de pós-processamento tem como objetivo a avaliação e a interpretação dos resultados em conjunto com especialistas de domínio.

A etapa de pré-processamento é a mais custosa desse processo e inclui a adequação dos dados para um formato no qual seja possível a aplicação de métodos para extração de padrões. Esses métodos necessitam de conjuntos de dados bem estruturados como, por exemplo, em forma de uma tabela atributo-valor. Na área médica, esses dados se apresentam em diversos formatos [Shortliffe and Barnett 2006] como imagens, formulários e laudos textuais. Esses formatos, no entanto, não são adequados para a aplicação de métodos de extração de padrões. Para que laudos textuais possam ser representados de forma adequada para aplicação de métodos computacionais de extração de padrões, seu conteúdo necessita ser mapeado em um formato atequado, como uma tabela atributo-valor. No entanto, o processo manual de mapeamento de laudos textuais em um formato adequado pode apresentar erros, além de ser lento e subjetivo [Lee 2005, Honorato et al. 2005, Cherman et al. 2008, Honorato et al. 2009b].

Para tonar mais rápido e efetivo e menos subjetivo o processo de mapeamento de laudos médicos textuais em um formato adequado podem ser utilizados métodos computacionais. É necessário então a construção de um algoritmo que seja capaz de compreender o que está representado nos termos descritos nas sentenças do laudo médico [Friedman and Johnson 2006, Lee 2005]. Para auxiliar na tarefa de compreensão dos termos presentes nos laudos médicos podemos utilizar uma estrutura específica para representação de conhecimento [Chute 2005, Revere and Fuller 2005, Bodenreider and Burgun 2005, Cherman et al. 2008, Honorato et al. 2009b]. Um exemplo desse tipo de estrutura é a ontologia a qual é uma representação formal dos conceitos de um determinado domínio e de seus relacionamentos [Gruber 1993, Carvalheira 2007].

O objetivo deste trabalho em andamento consiste em construir uma ontologia para auxiliar no mapeamento de laudos médicos de exames de Endoscopia Digestiva Alta — EDA. Este trabalho constitui parte do projeto Análise Inteligente de Dados, desenvolvido em parceira 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 Faculdade de Ciências Médicas da Universidade Estadual de Campinas — UNICAMP/Campinas.

Este trabalho está organizado da seguinte maneira: na Seção 2 são apresentados os métodos e as ferramentas utilizadas para o desenvolvimento deste trabalho; na Seção 3 são discutidas as principais motivações para o desenvolvimento deste trabalho e apresentados os resultados preliminares; na Seção 4 são apresentados a conclusão e os trabalhos futuros.

2. Material e Método

Na área médica, é comum encontrar informações na forma de laudos médicos descritos textualmente. Esses laudos normalmente são compostos por sentenças em que o médico descreve em linguagem natural as observações a respeito da saúde do paciente. O exame de EDA é um exame frequentemente descrito no formato de laudo textual. Na Figura 1, é apresentado um exemplo de laudo médico de EDA.

Para que as informações registradas nesses laudos possam ser analisadas para extração de conhecimento é estritamente necessário que elas estejam representadas em um formato estruturado que possa ser processado por métodos de extração de padrões. O mapeamento manual desses laudos em bases de dados estruturadas apresenta-se como

```
ESÔFAGO
- Mucosa de terço distal com presença de erosões, não confluentes
- Calibre e distensibilidade normais
- Motilidade normal
- TEG situada a aproximadamente 3,0 cm acima do pinçamento diafragmático
```

Figura 1. Exemplo de parte de um laudo de EDA utilizado

um método lento, além de apresentar um determinado grau de subjetividade, pois pode ser influenciado por fatores subjetivos dos que realizam essa tarefa. Para reduzir o tempo e a subjetividade do mapeamento desses dados foi desenvolvido um método de mapeamento de laudos médicos [Honorato et al. 2005, Honorato et al. 2008].

2.1. Mapeamento de Laudos Médicos

O método de mapeamento de laudos é dividido em duas fases conforme apresentado na Figura 2.

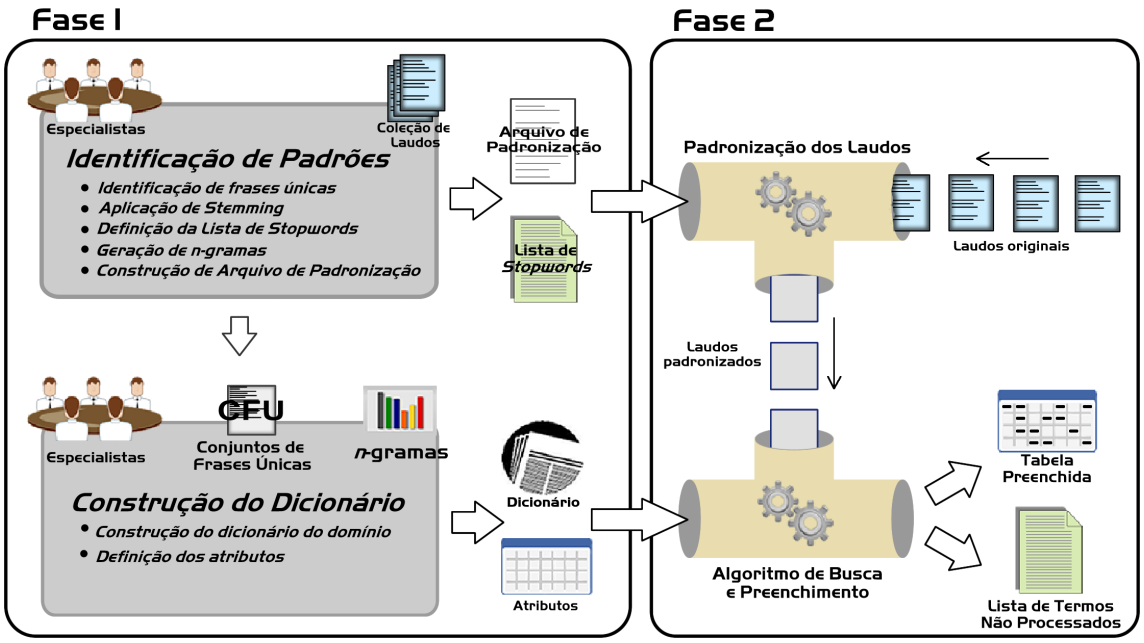

Figura 2. Método de mapeamento de laudos médicos para tabela atributo-valor

A primeira fase é dividida em duas etapas, realizadas com o auxílio de especialistas, as quais são a identificação de padrões e a construção do dicionário. A identificação de padrões é composta pelas seguintes tarefas:

Identificação de frases únicas: Nesta tarefa são identificadas todas as frases dos laudos e eliminadas as frases repetidas formando um Conjunto de Frases Únicas - CFU;

Aplicação de *Stemming*: Nesta tarefa é realizada a aplicação de *stemming* aos termos dos laudos a fim de reduzir o CFU;

Definição da lista de *stopwords*: Nesta tarefa é definida uma lista de palavras que serão filtradas dos laudos para reduzir o CFU. Essas palavras são definidas, em conjunto com especialistas, como sendo não importantes para o mapeamento das informações, como artigos, conjunções e preposições;

Construção do arquivo de padronização: Nesta tarefa é construído um arquivo de padronizações no qual são definidas palavras ou expressões que serão substituídas por outras equivalentes a fim de unificar a descrição dos termos;

Geração de *n*-gramas: Nesta tarefa é realizada a geração de *n*-gramas a fim de identificar as unidades terminológicas de maior frequência nos laudos.

A construção do dicionário é composta pelas seguintes tarefas:

Construção do dicionário de domínio: Nesta tarefa é construído o dicionário, o qual é uma hierarquia de locais no qual cada local possui uma lista de características e cada característica, por sua vez, pode possuir uma lista de subcaracterísticas. O local representa uma região anatômica e as características representam as possíveis anormalidades observadas no exame, as subcaracterísticas representam possíveis especificações das características. A cada sequência de local e característica ou local, característica e subcaracterística é vinculado um atributo assim como o valor que deve ser preenchido naquele atributo;

Definição da tabela atributo-valor: Nesta tarefa são definidos os atributos que irão compor a tabela atributo-valor e os possíveis valores que podem ser preenchidos para cada atributo da tabela, com base em interações com especialistas e com o CFU, o arquivo de padronização, a geração de *n*-gramas e aplicação de *stemming* da fase anterior.

Na segunda fase é realizada a padronização dos laudos originais conforme estabelecido no arquivo de padronização e na lista de *stopwords* construídos na primeira fase, a aplicação das padronizações acarreta na generalização de determinadas informações contidas nos laudos abstraindo parte do conteúdo. Os laudos padronizados são processados por um algoritmo de busca e preenchimento que preenche, com base no dicionário construído, a tabela de atributos definida na primeira fase e também gera uma lista contendo todos os termos não processados.

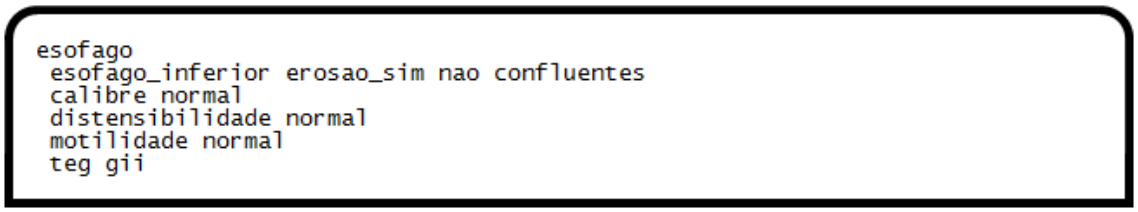

Figura 3. Exemplo de parte de um laudo de EDA padronizado

2.2. Construção de Ontologias

Uma ontologia é uma representação de um domínio de conhecimento específico, definida por meio de categorias (ou classes), propriedades (ou atributos) e relações dos indivíduos, os quais representam qualquer objeto existente no domínio.

Para a construção da ontologia foi utilizado o método descrito por [Uschold and Gruninger 1996] o qual é divido em três etapas principais:

Definição de escopo e objetivo: Nesta etapa define-se a especifidade da ontologia e o domínio a ser atendido pela ontologia. Definem-se todos os critérios os quais devem ser atendidos pela ontologia e também todas as ferramentas e a linguagem de representação as quais serão utilizadas na construção;

Construção: Nesta etapa constrói-se a ontologia através da definição dos conceitos que fazem parte do escopo definido na primeira etapa, assim como os relacionamentos existentes entre estes conceitos;

Avaliação: Nesta etapa avalia-se a expressividade e a consistência da ontologia a partir da especificação inicial dos critérios aos quais a ontologia deve atender. Caso algum critério não tenha sido satisfeito retorna-se à etapa de construção e realiza-se uma nova iteração de construção. Este ciclo prossegue até que todos os critérios sejam atendidos.

Para a construção foi utilizada a ferramenta Protégé [1] a qual é um dos editores de ontologias mais utilizados. As ontologias geradas pela ferramenta foram descritas em *Web Ontology Language* — OWL — , que é uma linguagem para representação de ontologias. Neste trabalho utiliza-se a sublinguagem OWL DL (Description Logic) que inclui todas as construções da OWL, porém com restrições de uso que garantem a completude e a computabilidade da ontologia [McGuinness and van Harmelen 2004].

Neste trabalho apresenta-se a construção da ontologia focada na seção do laudo correspondente ao esôfago.

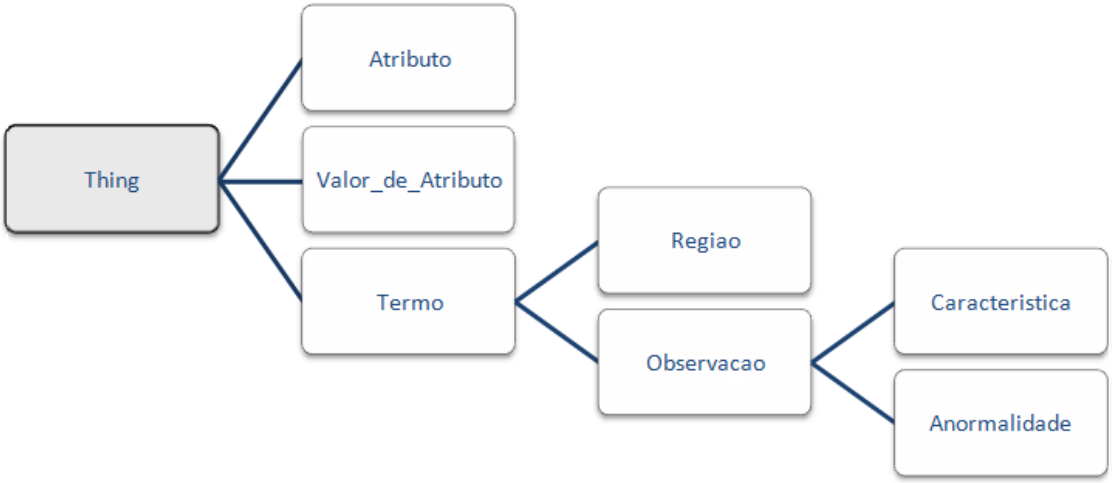

Figura 4. Representação geral dos conceitos da ontologia construída na primeira iteração

Os principais conceitos representados no dicionário, construído pelo método descrito anteriormente, foram determinados como classes, conforme apresentado na Figura 4. *Thing* é a classe principal que corresponde a todos os indivíduos, ela contém todas as outras classes. *Atributo* e *Valor_de_Atributo* são as classes diretamente envolvidas com a tabela atributo-valor, sendo a primeira correspondente a todos os atributos da tabela atributo-valor e o segundo correspondente a todos os valores possiveis para um atributo. Essas duas classes produzem subclasses que não são demonstradas na Figura 4, pois são apenas agrupamentos de atributos e valores de atributos afins. *Termo* corresponde aos termos presentes nos laudos padronizados, essa classe divide-se em dois tipos principais de termos: *Regiao* que corresponde aos termos que representam porções anatômicas e

[1] http://protege.stanford.edu/

Observacao que corresponde aos termos que caracterizam características (a classe *Caracteristica*) ou anormalidades (a classe *Anormalidade*) observadas pelo médico e descritas nos laudos.

2.3. Algoritmo de Mapeamento

Para utilizar a ontologia no mapeamento dos laudos padronizados para a tabela atributo-valor foi necessário também o desenvolvimento de um novo algoritmo de busca e preenchimento, que foi desenvolvido na linguagem de programação Java[2] com o auxílio de duas *Application Programming Interfaces* — API — para processamento da ontologia: o *Pellet Reasoner*[3] e a OWL API[4].

Esse algoritmo é divido em três etapas que são executadas para cada laudo:

Divisão do laudo em sentenças: Nesta etapa o laudo é dividido em um conjunto de sentenças;

Processamento das sentenças: Nesta etapa os termos são identificados e a cada termo é vinculada uma propriedade, conforme sua classificação definida na ontologia, e os termos não identificados são adicionados à uma lista de termos não processados. Em seguida o conjunto de propriedades mapeadas é associado segundo as regras de associação definidas e são identificados quais atributos devem ser preenchidos, assim como os valores que serão preenchidos. Essa etapa então é repetida em cada sentença do laudo até que todas tenham sido processadas. Cria-se então um conjunto de atributos e seus respectivos valores para esse conjunto de sentenças.

Preenchimento de atributos não-mapeados: Nesta etapa é consultada a ontologia para verificação dos atributos ausentes no conjunto de atributos mapeados, construído na etapa anterior, e também para identificação do valor padrão, conforme definido na ontologia, para cada um dos atributos que não foram mapeados.

Após a execução dessas três etapas para todos os laudos, obtêm-se uma tabela atributo-valor preenchida e uma lista de termos não mapeados. É importante ressaltar que o algoritmo é de complexidade linear em relação ao número total de termos a serem mapeados, isto é, o tempo de execução do algoritmo é linearmente dependente da quantidade de laudos a serem processados e da quantidade de termos a serem processados em cada laudo.

A nova abordagem foi aplicada a um conjunto de 609 laudos, realizados no período de 2000 a 2005 no Hospital Municipal de Paulínia. Esses laudos, descritos textualmente, estão divididos em quatro seções: esôfago; estômago; duodeno; observações e exames complementares. Como mencionado, neste trabalho, foram utilizadas apenas informações relacionadas ao esôfago.

3. Resultados Parciais e Discussão

Na fase 1 a identificação de padrões, realizada por meio da identificação do CFU, geração de *n*-gramas e intensa interação com especialistas, possibilitou a compreensão dos tipos de termos presentes nos laudos e como eles se relacionam entre si. Essa compreensão, por

[2]http://java.sun.com/

[3]http://clarkparsia.com/pellet/

[4]http://owlapi.sourceforge.net/

sua vez, é vital para a construção de uma ontologia correta. Desse modo, nessa mesma fase, foi construída a ontologia para auxiliar no mapeamento dos laudos médicos textuais de EDA para bases de dados estruturadas. É importante ressaltar que, de acordo com a pesquisa bibliográfica realizada, não foi encontrada ontologia, em língua portuguesa, construída para o tema tratado neste trabalho.

Na fase 2 foi utilizado o algoritmo desenvolvido para o mapeamento das informações presentes nos laudos em uma tabela atributo-valor com auxílio da ontologia construída na fase anterior. Foi mapeada a seção correspondente ao esôfago de cada laudo de EDA, do mesmo conjunto de 609 laudos, para uma tabela atributo-valor, a qual foi definida na fase anterior. Foram mapeados 100% dos atributos esperados para a seção de esôfago dos laudos alcançando a mesma taxa de precisão obtida utilizando o método baseado no dicionário.

No LABI têm sido desenvolvidos outros trabalhos no âmbito do mapeamento de laudos médicos textuais para bases de dados estruturadas, os quais objetivam a melhoria do método de mapeamento [Honorato et al. 2005, Honorato et al. 2008] e o desenvolvimento de ferramentas computacionais para dar suporte às etapas do método de mapeamento [Honorato et al. 2009a]. O método também tem sido aplicado, com bons resultados, em conjuntos de laudos médicos tanto de EDA como de Coloscopia [Cherman et al. 2007, Cherman et al. 2008].

Embora o dicionário utilizado pelo método tenha capacidade de mapear 100% dos valores esperados, na seção de esôfago, e a quase totalidade da informação contida nos laudos, sua classificação em locais, características e subcaracterísticas representa parcialmente a relação entre os termos.

A substituição do dicionário por uma ontologia possibilita uma melhor representação dos termos presentes nos laudos, pois os termos podem ser classificados de maneira mais específica. A utilização da ontologia também torna possível a adição de conteúdo semântico a cada um desses termos, de modo que o mapeamento seja mais efetivo e completo. Essa adição de conteúdo à estrutura pode tornar desnecessárias determinadas padronizações. Com a diminuição das padronizações menos conteúdo é abstraído dos laudos, de modo que o laudo padronizado apresente um conteúdo mais rico, possibilitando a adição de novos atributos à tabela atributo-valor e aumentando a quantidade de conteúdo mapeado dos laudos.

De acordo com os resultados apresentados, constatou-se que a nova abordagem do método de mapeamento, apresentada nesse trabalho, apresentou capacidade de mapeamento equivalente à anterior. Além disso, a representação por meio de uma ontologia permitiu uma melhor modelagem do conhecimento presente no laudo, permitindo descrever de modo mais completo as informações em relação ao modelo de relações hierárquicas entre locais, características e subcaracterísticas, proporcionando um maior refinamento na modelagem do domínio descrito nos laudos médicos. Essa primeira ontologia constitui a base de representação dos conceitos relacionados com o mapeamento de laudos de EDA da seção correspondente ao esôfago. A ontologia então será ampliada com o intuito de obter uma ontologia geral do domínio de EDA, a qual também será utilizada para o mapeamento completo dos laudos de EDA para bases de dados estruturadas, assim como para outras tarefas baseadas em conhecimento que possam ser aplicadas ao domínio de EDA,

como as etapas posteriores do processo de MD.

4. Conclusão e Trabalhos Futuros

Neste trabalho foi construída uma ontologia para auxiliar no processo de mapeamento de laudos textuais para bases de dados estruturadas. Primeiramente, o método de mapeamento de laudos médicos textuais para bases de dados estruturadas foi adaptado para utilizar a ontologia. Com a aplicação dessa nova abordagem foram mapeadas as seções correspondentes ao esôfago do conjunto de 609 laudos de EDA para uma tabela atributo-valor. Esse mapeamento foi realizado por meio de um algoritmo de busca e preenchimento baseado na ontologia construída, mapeando 100% dos atributos esperados.

O mapeamento desses laudos médicos para bases de dados estruturadas é importante para o processo de MD, pois é necessário que os dados estejam num formato adequado para aplicação de métodos de extração de padrões.

Os trabalhos futuros incluem: a finalização da ontologia atual para realizar o mapeamento de todo o laudo de EDA; a expansão da ontologia a fim permitir o mapeamento de mais informações, que são abstraídas no mapeamento atual, de modo a agregar mais informação à tabela atributo-valor; a adaptação do método baseado na ontologia para aplicação em outros domínios da área médica como em Coloscopia e Manometria Anorretal; a expansão para uma ontologia geral do domínio de EDA sendo útil para diversas tarefas baseadas em conhecimento que podem ser aplicadas a laudos de EDA, como as etapas subsequentes de MD.

5. Agradecimentos

Ao PTI Ciência e Tecnologia — PTI C&T — da Fundação Parque Tecnológico Itaipu — FPTI pelo auxílio na realização deste trabalho por meio da linha de financiamento de bolsas.

Referências

Bodenreider, O. and Burgun, A. (2005). Biomedical ontologies. In Chen, H., Fuller, S. S., Friedman, C., and Hersh, W., editors, *Medical Informatics*, chapter 8, pages 211–236. Springer.

Carvalheira, L. C. D. C. (2007). Método semi-automático de construção de ontologias parciais de domínio com base em textos. Master's thesis, Escola Politécnica da Universidade de São Paulo, São Paulo, SP.

Cherman, E. A., Lee, H. D., Honorato, D. D. F., Fagundes, J. J., Góes, J. R. N., Coy, C. S. R., and Wu, F. C. (2007). Metodologia de mapeamento de laudos médicos para bases de dados: Aplicação em laudos colonoscópicos. In *Anais do II Congresso da Academia Trinacional de Ciências (C3N)*, pages 1–9, Foz do Iguaçu, PR, Brasil.

Cherman, E. A., Spolaôr, N., Lee, H. D., Costa, L. H. D., Fagundes, J. J., Coy, C. S. R., and Wu, F. C. (2008). Metodologia de mapeamento computacional de informações médicas: Aplicação em laudos de coloscopia e manometria anorretal. *Revista Brasileira de Coloproctologia, 57º Congresso Brasileiro de Coloproctologia*, 29:42–42.

Chute, C. G. (2005). Medical concept representation. In Chen, H., Fuller, S. S., Friedman, C., and Hersh, W., editors, *Medical Informatics*, chapter 6, pages 163–182. Springer.

Friedman, C. and Johnson, S. B. (2006). Natural language and text processing in biomedicine. In Shortliffe, E. H. and Cimino, J. J., editors, *Biomedical Informatics*, chapter 2, pages 312–343. Springer.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. *Formal Ontology in Conceptual Analysis and Knowledge Representation*.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*, chapter 10.4 Text Mining. Morgan Kaufmann.

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

Honorato, D. D. F., Lee, H. D., Monard, M. C., Wu, F. C., Machado, R. B., Neto, A. P., and Ferrero, C. A. (2005). 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, Brasil.

Honorato, D. D. F., Monard, M. C., Lee, H. D., Ferrero, C. A., and Wu, F. C. (2009a). TP-Discover: um ambiente computacional para auxílio no pré-processamento de laudos médicos não-estruturados. In *Anais do XIII Simposio Informática y Salud (SIS), 38º Jornadas Argentinas de Informática (JAIIO), A ser publicado*, pages 1–10, Mar del Plata, BA, Argentina.

Honorato, D. D. F., Monard, M. C., Lee, H. D., Neto, A. P., and Wu, F. C. (2009b). Avaliação de um método de mapeamento de laudos médicos para uma representação estruturada: estudo de caso com laudos de endoscopia digestiva alta. In *Anais do IX Workshop de Informática Médica (WIM), XXIX Congresso da Sociedade Brasileira de Computação (CSBC)*, pages 1–10, Bento Gonçalves, RS, Brasil.

Lee, H. D. (2005). *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. Tese de doutorado, ICMC-USP, São Carlos, SP, Brasil.

McGuinness, D. L. and van Harmelen, F. (2004). *OWL Web Ontology Language Overview*. Disponível em: http://www.w3.org/TR/owl-features/. Acesso em: 20 jul 09.

Revere, D. and Fuller, S. S. (2005). Characterizing biomedical concept relationships. In Chen, H., Fuller, S. S., Friedman, C., and Hersh, W., editors, *Medical Informatics*, chapter 7, pages 183–210. Springer.

Rezende, S. O. (2003). *Sistemas Inteligentes - Fundamentos e Aplicações*. Manole, Barueri-SP, Brasil, 1 edition.

Shortliffe, E. H. and Barnett, G. O. (2006). Biomedical data: Their acquisition, storage, and use. In Shortliffe, E. H. and Cimino, J. J., editors, *Biomedical Informatics*, chapter 2, pages 46–79. Springer.

Uschold, M. and Gruninger, M. (1996). Ontologies: principles, methods and applications. *Knowledge Engineering Review*, 11:93–155.