Uma Metodologia para Estruturação de Laudos Médicos usando Ontologias*

Oscar Picchi Netto[1,2], Alessandra Alaniz Macedo[1], Paulo Mazzoncini de Azevedo Marques[2], José Augusto Baranauskas[1]

[1]Departamento de Computação e Matemática — DCM
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto — FFCLRP
Universidade de São Paulo — USP
Av. Bandeirantes, 3900 - Ribeirão Preto, SP - 14040-901 - Brasil

[2]Faculdade de Medicina de Ribeirão Preto — FMRP
Universidade de São Paulo — USP
Av. Bandeirantes, 3900 - Ribeirão Preto, SP - 14040-901 - Brasil

oscarpn@usp.br, ale.alaniz@usp.br, pmarques@fmrp.usp.br, augusto@usp.br

Abstract. *In computer systems aimed at generating and managing reports in radiology dominates the acquisition and storage of information in an open form of textual annotation. A major challenge faced in the development of these systems lies in the inadequacy of environments characterized by standardized interfaces that severely restrict the freedom of physicians in filling their reports, in contrast to the ineffectiveness of open text environments, which restrict a further analysis by computers. This paper aims to propose a methodology for structuring radiological reports from the Hospital das Clínicas da Faculdade de Medicina at Ribeirão Preto aiming in the future to build semi-automatically a report structure that allows knowledge extraction in order to facilitate the activities involved in teaching and researching in this hospital school. The methodology was evaluated on three bases each containing 5000 reports and the results compared with a standard ontology.*

Resumo. *Nos sistemas informatizados voltados à geração de laudos em radiologia predomina a aquisição e o armazenamento da informação na forma de anotação textual aberta. Um grande desafio enfrentado no desenvolvimento destes sistemas reside na inadequabilidade de ambientes caracterizados por interfaces de preenchimento padronizados e pré-definidos, que restringem fortemente a liberdade do médico na geração de seus relatos; contrastando com a ineficácia de outros ambientes que trabalham com textos abertos, restringindo fortemente as possibilidades de análise futura da informação. Este trabalho tem por objetivo propor uma metodologia de estruturação para a base de laudos radiológicos do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto visando, em trabalhos futuros, construir de forma (semi-)automática um laudo estruturado que permita a extração de conhecimento para facilitar as atividades ligadas ao ensino e pesquisa deste hospital escola. A metodologia proposta foi avaliada em três bases contendo 5000 laudos cada uma e os resultados comparados com uma ontologia padrão.*

*Projeto de pesquisa realizado com apoio financeiro FAPESP e CNPq/FAPEAM — INCT Adapta.

1. Introdução

Com o avanço da tecnologia nos últimos anos ficou muito fácil armazenar grandes quantidades de dados. Na área médica, esse crescimento é notável dada toda variedade de informações de pacientes que se encontram em formulários médicos, processos laboratoriais e laudos médicos armazenados atualmente sob a forma digital. Especificamente, no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HCFMRP) todos os dias centenas de exames radiológicos são realizados. Em muitos destes exames são geradas milhares de imagens, as quais são então utilizadas pelos especialistas médicos para gerar um laudo para cada exame no formato de texto livre. Tal laudo, junto com as imagens e informações complementares sobre o paciente (idade, sexo, exames realizados, etc), permite o diagnóstico médico dos pacientes. Portanto, laudos de radiologia contêm uma grande quantidade de informação que caracteriza a condição médica do paciente. Todavia, uma grande porcentagem desta informação não é estruturada, assumindo a forma de texto livre o que dificulta processos computacionais de busca, ordenação e análise.

Estudos na literatura mostram os benefícios potenciais de informações médicas estruturadas envolvendo práticas médicas, pesquisa e ensino. Na prática clínica, relatórios estruturados podem auxiliar na organização e melhoria da apresentação de registros médicos [Shortliffe and Hubbard 1989, Aberle et al. 1996]. Na pesquisa e ensino, relatórios estruturados podem melhorar significativamente a precisão e relevância em tarefas de recuperação de informação. Apenas dados estruturados viabilizam técnicas de modelagem de base de dados causal, espacial, temporal e evolucionária que estão em desenvolvimento nas áreas de informática médica e ciência da computação. Por exemplo, no estudo de [Honorato et al. 2009] é apresentada uma metodologia que pode ser aplicada 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. Os resultados encontrados fornecem um indicativo que a metodologia pode auxiliar na redução do tempo de atuação dos especialistas na análise de grandes quantidades de documentos não estruturados. Já no estudo de [Taira et al. 2001] é desenvolvida uma abordagem para um processador de linguagem natural que automaticamente estrutura informações médicas importantes que estão contidas em um documento de radiologia em formato de texto livre. Este sistema não necessita de nenhuma ajuda de especialistas no domínio, utilizando extensivamente métodos estatísticos e de Aprendizado de Máquina.

Neste contexto, quando as informações encontram-se em forma estruturada, como em uma base de dados, é possível realizar a busca por modelos para suporte à tomada de decisão médica por meio de algoritmos de Aprendizado de Máquina [Mitchell 1997]. Uma vez que os textos dos laudos não estão estruturados é necessário um pré-processamento dos textos. O pré-processamento deve ser adequado, de forma a permitir que os modelos extraídos representem adequadamente o conhecimento embutido nos textos. Em geral, o pré-processamento é uma etapa de um processo mais amplo, denominado Mineração de Textos e consiste em três etapas elementares [Feldman and Sanger 2006]: (i) pré-processamento, (ii) mineração e (iii) pós-processamento. Na etapa de pré-processamento ocorre a preparação dos dados (para o processo de mineração de textos) na qual o texto (formato não estruturado) é transformado em alguma representação estruturada. Existem diversas formas de representação sendo a mais utilizada o formato de uma tabela atributo-valor: em geral, as palavras existentes nos textos são transformadas em atributos

com sua frequência associada. Esta tabela é então utilizada na etapa de mineração, na qual, em geral, são utilizados algoritmos de Aprendizado de Máquina. No Aprendizado de Máquina utiliza-se inferência indutiva a partir de um conjunto de exemplos), induzindo um modelo sobre o conceito intrínseco nos dados. Na etapa de pós-processamento ocorre a interpretação dos resultados, que consiste na avaliação dos modelos extraídos. Assim como o armazenamento de grandes volumes de dados é de pouca valia, a menos que existam métodos computacionais adequados para analisá-los, também é pouco produtivo extrair modelos simbólicos que não causem surpresa ou representem conhecimento novo ou mesmo que sejam altamente redundantes para o especialista do domínio. É nesta etapa que ocorre a remoção de padrões irrelevantes ou redundantes e tradução de padrões úteis em termos inteligíveis pelos especialistas. Também é nessa etapa que ocorre o uso do conhecimento extraído, que consiste na incorporação do conhecimento ao domínio, seja tomando ações baseadas no conhecimento novo ou simplesmente documentando e relatando para as partes interessadas o conhecimento obtido, bem como remoção de conflitos potenciais com conhecimento previamente tido como correto.

Neste artigo é proposta uma metodologia de extração de termos em laudos médicos de forma estruturada. A proposta é que em trabalhos futuros e complementares seja possível construir de maneira (semi-)automática um laudo estruturado de forma a facilitar a extração de conhecimento, além de facilitar atividades ligadas ao ensino e pesquisa, uma vez que o estudo é em um hospital escola. A metodologia proposta foi avaliada em três bases laudos e os resultados comparados com uma ontologia padrão.

O restante deste trabalho está organizado da seguinte maneira: na Seção 2 é feita uma introdução sobre laudos médicos. Na Seção 3 é apresentada a metodologia proposta para a estruturação de laudos. Experimentos e resultados obtidos com uma base de laudos são apresentados na Seção 4 e, finalmente, na Seção 5 são apresentadas as conclusões.

2. Laudo Médico

O termo *laudo médico* é comumente utilizado para a interpretação de exame complementar ou resultado de perícia médica elaborado por médico. Na área médica, laudos são comumente descritos na forma textual, normalmente compostos por sentenças em que o médico descreve em linguagem natural as observações a respeito da saúde do paciente. Normalmente, o médico responsável por fazer o exame escreve suas suspeitas de acordo com o exame realizado e as imagens obtidas deste exame. Outro médico utilizando as mesmas imagens valida este laudo, confirmando assim o mesmo. O laudo então é levado ao médico que requisitou o exame e este, por meio do laudo do paciente e das informações previamente obtidas com o paciente, fornece o diagnóstico.

O HCFMRP possui um sistema de laudo digitalizado na realização de exames radiológicos, armazenando seus resultados na tabela Exame Radiológico do seu banco de dados. Nessa tabela existem diversos campos estruturados tais como código do paciente, número do exame, código da região anatômica, data da realização do laudo bem como campos no formato de texto livre tais como descrição e conclusão, dentre outros. São nestes campos expressos em linguagem natural que a mineração de textos mostra-se adequada para extrair ou estruturar o conhecimento, como o exemplo exibido na Figura 1. Para tanto, é necessário que os textos sejam representados em um formato estruturado que possa ser processado por algoritmos de extração de conhecimento. Uma forma diferente

Descrição: 'Controle pós-operatório de segmentectomia evidencia: Reduçao das dimensoes do lobo direito do fígado, com hipertrofia compensatória do lobo esquerdo. Baço pâncreas e rins com forma, contornos, dimensões e intensidade de sinal normais. Dilataçao de vias biliares intra-hepática e ductos hepáticos direito e esquerdo. O ducto hepático comum não é identificado e do ducto biliar comum (colédoco), visualiza-se apenas o terço distal que apresenta calibre normal. Na região do segmento IV junto à cúpula frênica, observa-se ductos biliares com dilatações saculares , com pequena coleção subcapsular com hipersinal em T2 (bilioma?). Vesícula biliar nao visualizada. Alças intestinais e gordura peritoneal e retroperitoneal sem alteração'

Conclusão: 'Dilatação de vias biliares intra hepáticas (Obstrução ao nível do hepático comum?), com formação de pequena coleção (bilioma) junto à cúpula frênica.'

Figura 1. Exemplo dos campos de *descrição* e *conclusão* de um laudo radiológico

daquela adotada neste estudo consistiria no mapeamento manual desses laudos em bases de dados estruturadas, o que apresenta-se como um método lento e com um determinado grau de subjetividade, uma vez que pode ser influenciado por fatores específicos do ser humano que realiza esta tarefa.

3. Metodologia Proposta

Existem várias maneiras de transformar textos em dados estruturados. Uma delas é transformá-los em uma representação atributo-valor utilizando a abordagem *bag of words*, na qual a frequência das palavras (termos), independente de seu contexto ou significado, são contadas. A ocorrência de palavras em sequência, em geral, pode conter mais informação do que palavras ocorrendo isoladamente. Portanto, ao criar atributos pela junção de duas ou mais palavras consecutivas, é esperado obter atributos com um maior poder de predição, compondo assim o n-grama, onde n representa o número de palavras que foram unidas para a geração de um atributo. Por exemplo, considerar as palavras *lobo* e *direito* individualmente pode agregar pouco conhecimento, pois *lobo* pode referir-se a um *animal* ou *região anatômica* e *direito* pode ser *uma faculdade de praticar algo* ou *oposto de esquerdo*, dentre outros significados. Todavia, o termo composto *lobo direito* pode agregar muito mais informação se o texto se refere a uma região anatômica do corpo humano.

A partir da contagem dos n-gramas é possível gerar uma tabela cujas entradas contêm informações relacionadas à frequência de cada n-grama. Neste estudo foi utilizada a ferramenta PreText [Soares et al. 2008] que permite a geração de n-gramas para qualquer valor de n. A partir dos conjuntos n-gramas é possível aplicar os demais passos do Algoritmo 1 para geração de uma ontologia.

Uma ontologia é especificada por uma coleção de termos (ou conceitos) e seus relacionamentos definindo uma ordem parcial, em geral, da forma *tipo-subtipo* [Gruber 1993, Sowa 2000]. Formalmente, dada uma linguagem lógica L, onde $L_p \subset L$ é o conjunto de símbolos predicados de L e $L_f \subset L$ é o conjunto de fórmulas bem formadas de L, uma ontologia é uma tupla (V, A) na qual o vocabulário $V \subset L_p$ e os axiomas $A \subset L_f$ [Heflin et al. 1999]. Taxonomias que especificam relacionamentos hierárquicos entre termos de um domínio específico do conhecimento humano — representados na forma de árvores ou grafos direcionados acíclicos — estão entre as ontologias mais utilizadas [Zhang et al. 2002]. Nesse caso, cada conceito tem a ele uma ontologia na forma de uma hierarquia (árvore), onde cada aresta da árvore representa um relacionamento entre entre dois nós. Os relacionamentos podem ser do tipo *é-um*, *é-parte-de* ou *é-um-processo-de*, dentre outros.

Algoritmo 1 Algoritmo para criação da ontologia a partir de laudos

Require: Documentos, uma coleção de documentos contendo laudos radiológicos
Região, uma região anatômica de interesse
N, número máximo de N-gramas a serem gerados (default 10)
Ensure: Ontologia, uma ontologia criada a partir dos Documentos
1: Docs_Selecionados ← Selecione(Documentos, Região)
2: LC ← ∅
3: $NDocs$ ← Número_de_Documentos(Docs_Selecionados)
4: (1-grama, 2-grama, ..., N-grama) ← PreText(Docs_Selecionados)
5: **for** i ← 1 **to** Número_de_Palavras(1-gram) **do**
6: **if** Frequência(palavra$_i$) $> \theta \times NDocs$ **then**
7: LC ← LC ∪ {palavra$_i$}
8: **end if**
9: **end for**
10: Ontologia ← LC
11: **for** $n \leftarrow 2$ **to** N **do**
12: **for** $i \leftarrow 1$ **to** Número_de_ Palavras(LC) **do**
13: **if** $palavra_i \subset n$-grama **then**
14: Associe $palavra_i$ ao nó correspondente na Ontologia
15: **end if**
16: **end for**
17: **end for**
18: **return** Ontologia

O Algoritmo 1 utiliza um conjunto de documentos e a região associada a esses documentos. Todas as palavras contidas no conjunto 1-grama são candidatas a estarem na LC (lista de candidatas). O algoritmo faz uso do parâmetro θ, onde $0 \leq \theta \leq 1$, explicado a seguir. Se $\theta = 0$ então todas as palavras, independente da sua frequência são adicionadas à LC; caso $\theta > 0$ as palavras do conjunto 1-grama com frequência maior que $\theta \times NDocs$, onde $NDocs$ é o número de documentos selecionados, serão adicionadas à LC (linhas 3 à 9). Esta estratégia foi adotada pois, em geral, n-gramas que apresentam uma maior frequência podem ser mais representativos para o aprendizado. Após isso, para todos os conjuntos n-gramas ($n = 2, 3, \ldots, N$) e para todas as palavras contidas na LC, o algoritmo verifica se a palavra$_i$ da LC está contida na lista de palavras do conjunto n-grama; em caso afirmativo associa a palavra$_i$ ao nó correspondente na Ontologia (linhas 10 à 17).

Por exemplo, suponha uma base de dados radiológicos hipotética contendo apenas exames de joelho. Primeiramente, os documentos são processados pelo PreText, gerando assim os conjuntos n-gramas. Neste exemplo, assuma que $N = 3$, ou seja, que serão gerados os conjuntos 1-grama, 2-gramas e 3-gramas. Agora suponha que no conjunto 1-grama aparece a palavra *joelho* e que a frequência dela seja maior do que $\theta \times NDocs$. Sendo assim *joelho* entra na LC. Assuma então que no conjunto 2-gramas aparecem as seguintes palavras *joelho_ligamento* e *joelho_menisco* e que no conjunto 3-gramas existam *joelho_ligamento_lateral*, *joelho_ligamento_posterior* e *joelho_ligamento_cruzado*. Então *joelho_ligamento* e *joelho_menisco* são colocados no sub-ramo cuja raiz é *joelho*. Analogamente, *joelho_ligamento_lateral*, *joelho_ligamento_posterior* e *joelho_ligamento_cruzado* são colocados no sub-ramo cuja raiz é *joelho_ligamento*. Com essas suposições é mostrada a ontologia gerada pelo Algoritmo 1 na Figura 2.

4. Experimentos, Resultados & Discussão

Para avaliar a metodologia proposta neste estudo foram utilizados laudos radiológicos do base de dados do HCFMRP, atualizada até a data de 05/02/2009. Os dados consistem em exames realizados pelo Centro de Ciências das Imagens e Física Médica e inseridos no

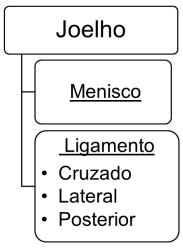

Figura 2. Exemplo de ontologia criada sobre laudos hipotéticos de joelho

sistema desde 1999 até fevereiro de 2009. A pesquisa foi aprovada pelo Comitê de Ética em Pesquisa do HCFMRP, sob processo número 10791/2007. Essa base possui 1.134.071 laudos, porém nem todos possuem todos os campos mencionados na Seção 2. Portanto foi necessário pré-processar esta base de dados para utilizar somente laudos completos; esta redução da base de dados foi de 570.641 laudos, restando assim 563.430 laudos completos para serem analisados. A base de laudos completos encontra-se sem nenhum pré-processamento adicional, sendo composta por todos os dados da tabela Exame Radiológico do banco de dados relacional do HCFMRP. Como é possível notar, no Algoritmo 1 é possível selecionar uma região anatômica específica, como joelho, pulmão, mama ou até mesmo região não tão específica como tórax e abdômen.

Portanto, com base no Algoritmo 1 foram criadas 3 ontologias, utilizando 3 bases de 5000 laudos cada, selecionados de forma aleatória dentre os 563.430 laudos completos: uma base mais específica (joelho), uma base intermediária (tórax) e uma base geral (utilizando todos os tipos de laudo). Em cada uma dessas bases foram realizados experimentos para a geração das ontologias, utilizando-se $\theta = 0,045$, valor definido de forma heurística. É importante salientar que a ferramenta PreText utilizada trabalha com 'radicais' (*stems*) das palavras. Assim sendo, nesta fase inicial da metodologia proposta, as ontologias aqui mostradas contém apenas os radicais encontrados pelo PreText.

De forma a avaliar a metodologia proposta seria possível (i) utilizar o auxílio de especialistas médicos no domínio da aplicação e/ou (ii) utilizar uma ontologia pré-estabelecida e confiável e procurar por partes que sejam comuns àquelas encontradas pela metodologia aqui proposta, dentre outras formas [Brank et al. 2007]. Neste estudo efetuou-se uma avaliação intermediária entre (i) e (ii), comparando-se manualmente os resultados obtidos com uma ontologia internacionalmente conhecida e validada, a saber RadLex[1], procurando por ramos que fossem comuns a ambas ontologias. Há, entretanto, o obstáculo para se fazer comparações pelo fato que a ontologia RadLex está escrita na língua inglesa, enquanto que os laudos utilizados neste estudo estão na língua portuguesa, o que impede a comparação automática de ontologias, por exemplo, pelo algoritmo proposto por [Brank et al. 2007]. Todavia, é importante salientar que é de igualmente importante a validação por especialistas, algo que poderá ser efetuado em trabalhos futuros.

Um resumo sobre os experimentos realizados pode ser encontrado na Tabela 1. Nesta tabela são mostrados os parâmetros utilizados bem como o número de palavras e de *stems* de cada base e o número de palavras no primeiro LC(1) e no segundo LC(2) níveis da ontologia. Um fato que merece atenção é que diante do grande número de palavras e

[1] http://radlex.org/viewer

radicais (*stems*) a metodologia proposta gerou poucos candidatos para o primeiros níveis da ontologia, o que é interessante do ponto de vista de estruturação do conhecimento. A seguir, são descritos os resultados em cada uma das 3 bases de laudos selecionados para estudo. As ontologias criadas foram visualizadas utilizando o software Protégé[2].

Tabela 1. Resumo sobre os experimentos

Base	θ	N	Documentos	Palavras	Stems	LC(1)	LC(2)
Joelho	$0,045$	10	5000	172.145	5.880	1	20
Tórax	$0,045$	10	5000	154.413	10.665	1	10
Geral	$0,045$	10	5000	312.008	17.514	2	29

4.1. Base Específica - Joelho

Neste experimento a base de laudos selecionada contém apenas laudos referentes a região anatômica do joelho. Após a aplicação do Algoritmo 1, obtém-se a ontologia mostrada à esquerda da Figura 3. Neste experimento utilizando uma base bem específica foi possível encontrar dois sub-ramos cuja a representação pode ser encontrada no RadLex de forma quase total, ou seja, quase todas as folhas selecionadas aparecem no RadLex. Os sub-ramos de *ligament* (radical de *ligamento*) e *metafis* (radical de *metáfise*) gerados na ontologia foram encontrados no RadLex. Nas Figuras 4(a) e 5(a) são mostrados os resultados encontrados pela metodologia aqui proposta; nas Figuras 4(b) e 5(b) são mostradas partes da ontologia que foram encontradas no RadLex; os nomes sublinhados aparecem na ontologia gerada pelo Algoritmo 1.

4.2. Base Intermediaria - Tórax

Neste experimento utilizou-se uma base de laudos intermediária, ou seja, selecionando uma região anatômica e, portanto, vários tipos de exames e, consequentemente, vários tipos de laudos, aumentando um pouco o escopo em relação ao experimento anterior. Observa-se que, apesar de ser uma base mais geral, o número de palavras diminiu; isso ocorreu pois alguns laudos de joelho são extremamente detalhados, enquanto que, ao que tudo indica, alguns laudos da região torácica não parecem apresentar um detalhamento tão profundo. Entretanto, isso requer a análise por especialistas no domínio para sua efetiva comprovação. Outro fato que pode ser observado é que o número de *stems* aumentou, mostrando que são necessários mais radicais de palavras para descrever uma região anatômica maior. Já na LC ocorreu a diminuição de 45 (joelho) para 24 (tórax). A ontologia gerada para esta base é mostrada no centro da Figura 3.

Para esta base de laudos não houve sub-ramos que possam ser encontrados no RadLex. Isso pode ter ocorrido por 2 motivos: o primeiro é que a base utilizada pode não representar uma amostra significativa da base original; o segundo motivo pode ser devido ao fato de que o algoritmo proposto não conseguiu trabalhar com uma LC tão reduzida a ponto de gerar algo que poderia ser encontrado no RadLex. Apesar disso, um dos sub-ramos que pode ser destacado por apresentar alguns resultados possivelmente relevantes. O radical *process* possui quatro filhos, todos esses filhos de *process* podem ser reconhecidos como processos que ocorrem no corpo humano como, por exemplo, o filho *inflamatori* que representa o *processo inflamatório*. Depois de uma busca no RadLex sobre a palavra inflamatório, observou-se que este não possui nenhuma menção a este

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

Figura 3. Ontologia criada sobre laudos de joelho (esquerda), laudos de tórax (centro) e laudos gerais (direita)

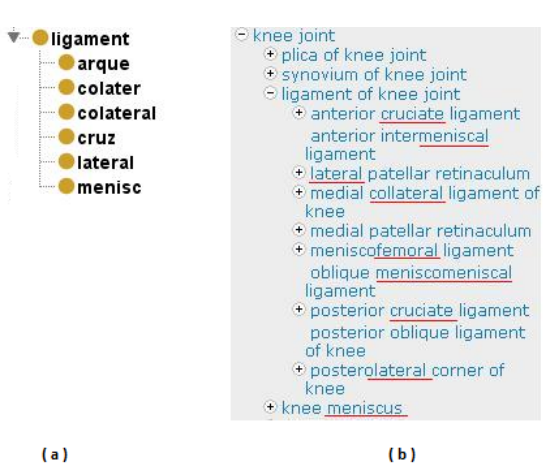

Figura 4. Subramo da ontologia Ligament (a) e sua contrapartida no RadLex (b)

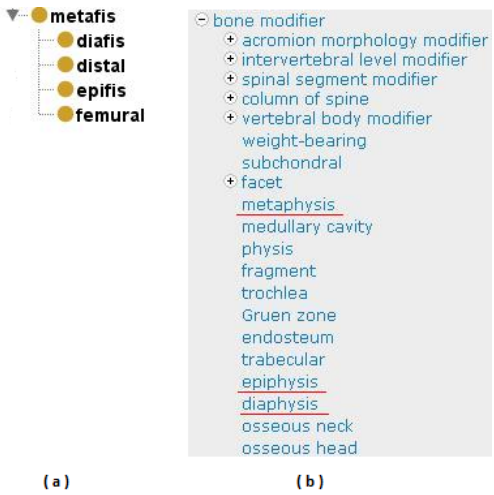

Figura 5. Subramo da ontologia Metafis (a) e sua contrapartida no RadLex (b)

processo. Nesta base, não foram encontrados resultados que possam ser achados em contrapartida no RadLex.

4.3. Base Geral

Neste experimento não houve nenhum tipo de restrição quanto a região anatômica. Neste caso, a base possui 312.008 palavras e 17.514 *stems*. É possível perceber que estes números são maiores que os das bases utilizadas anteriormente. A LC gerada para esta base possui 61 palavras. A ontologia gerada é mostrada à direita da Figura 3. Apesar de possuir uma LC maior que as das outras bases, a ontologia gerou um resultado que pode ser interessante: o sub-ramo *estenos*, radical da palavra *estenose*. Apesar de o termo *estenose* poder ser encontrado no RadLex, os filhos gerados pela ontologia não são encontrados no mesmo. Na Figura 6(a) e 6(b) são mostrados os dois sub-ramos, o da ontologia gerada pelo Algoritmo 1 e a ontologia do RadLex, respectivamente. Apesar de ser uma base mais geral que a base de tórax, esta base apresentou um dos sub-ramos presentes no RadLex, portanto, é possível observar que apesar de utilizar uma base geral o algoritmo é capaz de encontrar resultados que podem ser relevantes.

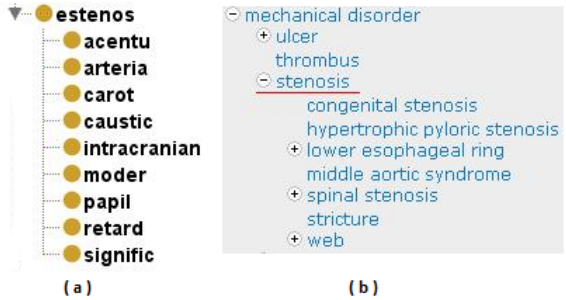

Figura 6. Subramo da ontologia *estenose*(a) e sua contrapartida no RadLex(b)

5. Conclusão

Nos sistemas informatizados voltados à geração de laudos em radiologia, predomina-se a aquisição e o armazenamento da informação na forma de anotação textual aberta. Um grande desafio enfrentado no desenvolvimento destes sistemas reside na inadequabilidade de ambientes caracterizados por interfaces de preenchimento padronizados e pré-definidos, que restringem fortemente a liberdade do médico na geração de seus relatos; contrastando com a ineficácia de outros ambientes que trabalham com textos abertos, restringindo fortemente as possibilidades de análise futura da informação.

Neste trabalho foi proposta uma metodologia que gera uma lista de candidatos a serem os primeiros sub-ramos de uma ontologia utilizando-se das palavras geradas e a frequência das palavras como métrica de corte. Foram geradas 3 ontologias, cada uma avaliando o comportamento da metodologia proposta em bases específica, intermediária e geral. Contatou-se que para uma base de 5000 laudos, os resultados foram mais interessantes para a base mais específica. Trabalhos adicionais podem utilizar mais níveis na ontologia, outros valores para a métrica de corte, além de utilizar um número de laudos maior. As ontologias também poderão ser mostradas com as palavras originais encontradas nos documentos ou mesmo agrupá-las em um nó da ontologia ao invés de somente

os radicais encontrados. Por último, seria importante para o país a adoção de ontologias padronizadas em língua portuguesa, o que permitiria avaliar mais objetivamente o conhecimento extraído por técnicas automáticas como a aqui proposta.

Agradecimentos

Este projeto de pesquisa foi desenvolvido na Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (USP), Faculdade de Medicina de Ribeirão Preto (USP) e Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto com apoio financeiro FAPESP, CNPq, FAPEAM e INCT Adapta.

References

Aberle, D. R., Dionisio, J. D., McNitt-Gray, M. F., Taira, R. K., Cárdenas, A. F., Goldin, J. G., Brown, K., Figlin, R., and Chu, W. W. (1996). Integrated multimedia timeline of medical images and data for thoracic oncology patients. *Radio-Graphics*, 16:669–681.

Brank, J., Grobelnik, M., and Mladenić, D. (2007). *Automatic Evaluation of Ontologies*, chapter 11. In [Kao and Poteet 2007].

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

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Heflin, J., Hendler, J., and Luke, S. (1999). Coping with changing ontologies in a distributed environment. In *Proceedings of AAAI-99 Workshop on Ontology Management*, pages 74–79. AAAI Press.

Honorato, D. F., Monard, M. C., Lee, H. D., Neto, A. P., and Chung, W. F. (2009). Avaliação de uma metodologia de mapaemento de laudos médicos para uma representação estruturada: estudo de caso com laudos de endoscopia digestiva alta. *WIM - IX Workshop de Informática Médica.*

Kao, A. and Poteet, S. R., editors (2007). *Natural Language Processing and Text Mining.* Springer.

Mitchell, T. M. (1997). *Machine Learning.* McGraw–Hill.

Shortliffe, E. H. and Hubbard, S. M. (1989). *Information systems in oncology. In: De Vita VT, Hellman S, Rosenberg S, eds. Cancer: principles and practice of oncology.*

Soares, M. V., Prati, R. C., and Monard, M. C. (2008). Pretext ii: Descrição da reestruturacão da ferramenta de pré-processamento de textos. Technical Report 333, ICMC-USP, São Carlos - SP.

Sowa, J. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Brooks Cole Publishing Co., Pacific Grove, CA.

Taira, R. K., Soderland, S. G., and Jakobovits, R. M. (2001). Automatic structuring of radiology free-text reports. *RadioGraphics*, 21:237–245.

Zhang, J., Silvescu, A., and Honavar, V. G. (2002). Ontology-driven induction of decision trees at multiple levels of abstraction. In *Proceedings of the 5th International Symposium on Abstraction, Reformulation and Approximation*, pages 316–323, London, UK. Springer-Verlag.