Representação de informações não-estruturadas de laudos de exames radiológicos utilizando modelagem interativa de mapas conceituais

Lucio Geronimo Valentin

TESE APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
DOUTOR EM CIÊNCIA DA COMPUTAÇÃO

Programa: Doutorado em Ciência da Computação
Orientador: Prof. Dr. Marcel Parolin Jackowski

São Paulo, agosto de 2016.

Representação de informações não-estruturadas de laudos de exames radiológicos utilizando modelagem interativa de mapas conceituais

Esta é a versão original da tese elaborada pelo candidato Lucio Geronimo Valentin.

Resumo

VALENTIN, L. G. Representação de informações não-estruturadas de laudos de exames radiológicos utilizando modelagem interativa de mapas conceituais. 2016. 76 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

A representação e subsequente extração de informações a partir de laudos de exames radiológicos permitem a obtenção de um retrato da saúde da população que realiza estes exames. Embora exista um grande esforço da comunidade médica para a estruturação textual de tais laudos, em sua maioria são confeccionados sem estruturação e podem não ser necessariamente completos no seu sentido linguístico. Como consequência, a extração e representação de informações se tornam tarefas não triviais para a normalização e modelagem do conhecimento armazenado. Esta dissertação apresenta uma abordagem para a representação de informações não estruturadas de laudos a partir de mapas conceituais. Os mapas são construídos de forma interativa por especialistas. Um processo de união dos mapas é utilizado para a criação de um modelo que é usado para identificar as mesmas relações em outros laudos. O principal diferencial desta metodologia é o baixo acoplamento de recursos semânticos específicos para um domínio e sua facilidade para a definição das regras de extração. Como resultado de implementação deste trabalho foram criados dois sistemas, um para extração de informações, denominado Miner@ e outro para a representação de informações, chamado de Analaudos. Os resultados iniciais mostraram um grau de precisão e cobertura. Com o desenvolvimento desta metodologia pretende-se facilitar a extração de informações de laudos e a criação de uma rica fonte de conhecimento para geração de estatísticas populacionais, para suporte à tomada de decisão médica e apoio na definição de políticas públicas de saúde, utilizando o enorme volume de laudos disponíveis atualmente em várias instituições públicas e privadas.

Palavras-chave: Mineração de textos, Ontologia, PLN, Laudos Médicos.

Abstract

VALENTIN, L. G. Representation of unstructured information of radiological reports using interactive modeling of conceptual maps. 2016. 76 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

The representation and subsequent extraction of information from radiological reports allow you to obtain a picture of the health of the population that performs these tests. Although there is a great effort from the medical community for the textual structure of such reports, most are made without structuring and may not necessarily be complete in its linguistic sense. As a result, extraction and representation of information become non-trivial tasks for the normalization and modeling the stored knowledge. This paper presents an approach to the representation of unstructured information reports from conceptual maps. The maps are built interactively by experts. A union process of the maps is used to create a model that is used to identify the same relations in other reports. The main distinguishing feature of this method is the low coupling specific semantic features for a domain and its ease for the definition of extraction rules. As a result of implementation of this work it was created two systems, one for extraction of information, called Miner@® and another for the representation of information, called Analaudos. Initial results showed high precision and recall. With the development of this methodology is intended to facilitate the extraction of information from reports and creating a rich source of knowledge for the generation of population statistics, to support medical decision-making and support the definition of public health policies, using the huge volume of reports currently available in several public and private institutions.

Palavras-chave: Text Mining, Ontology, NLP, Radiological Reports.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Os exames médicos baseados em imagens são um eficiente recurso para o diagnóstico, prognóstico e acompanhamento de patologias. Sua utilização permite a representação visual *in vivo* de órgãos e tecidos do corpo humano. Ultrassonografia, radiografia, ressonância magnética e tomografia são exemplos de modalidades de exames, que podem ser não invasivos ou invasivos quando utilizam algum contraste (Zivadinov et al., 2013).

O número de exames e o número de imagens por exame têm crescido nas últimas décadas no Brasil, como mostra a Figura 1 compilada a partir dos dados disponíveis em DATASUS-SIAB (2016). Somente no Sistema Único de Saúde (SUS), o número de exames de radiodiagnóstico realizados por ano passaram de 1,1 milhões em 2001 para 4,5 milhões em 2010 e para 5,2 milhões em 2015. Somente os exames de ultrassonografia passaram, no mesmo período, de 727 mil para 2,4 milhões e para 3,4 milhões. Dentre outros fatores, esse crescimento se deve a uma maior acessibilidade por parte da população a estes equipamentos. A Figura 2 mostra o crescimento ocorrido na última década no número de equipamento de diagnóstico por imagem disponíveis somente no estado de São Paulo. Os equipamentos de ultrassonografia e de raio-x são predominantemente mais acessíveis à população. Freitas e Yoshimura (2005) destacam que esta situação também é observada nos Estados Unidos, onde o número per capita de equipamentos é ainda maior.

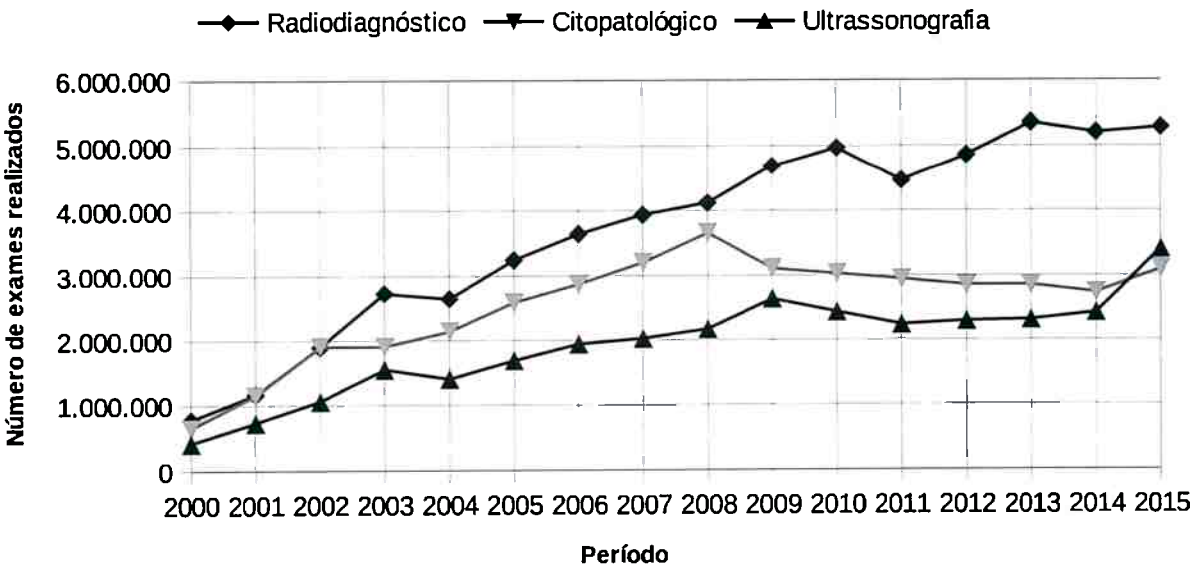

Figura 1: Exames realizados no Brasil entre 2000 e 2015. Dados obtidos em DATASUS-SIAB(2016).

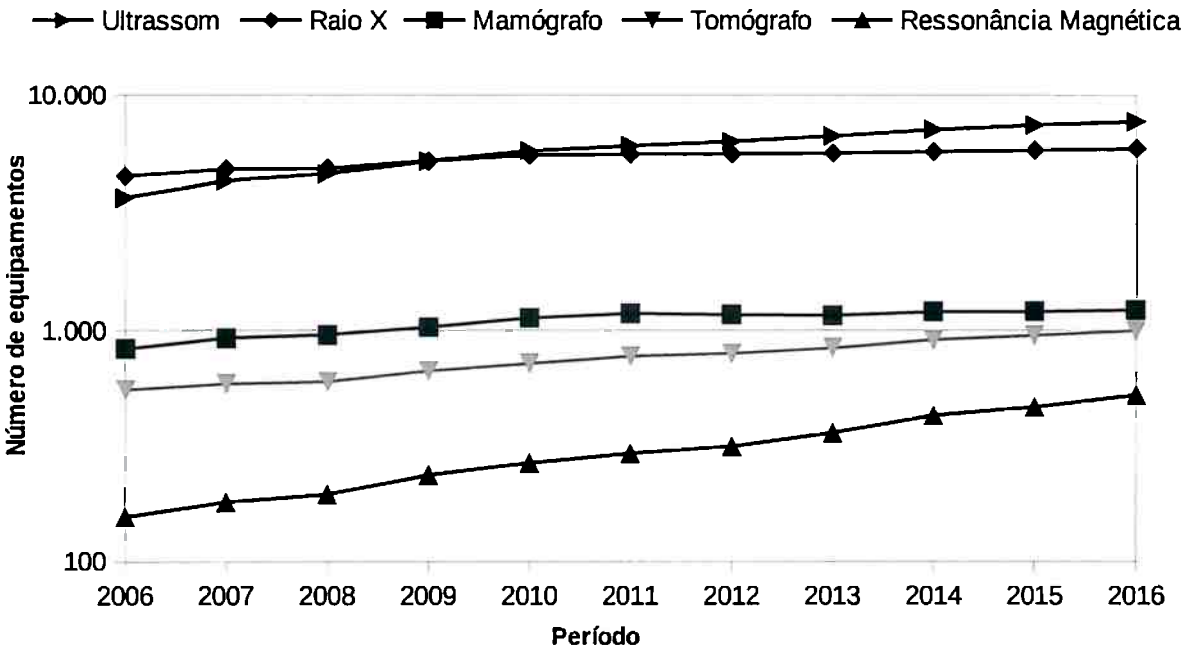

Figura 2: Número, em escala logarítmica, de equipamentos de diagnóstico por imagem disponíveis no estado de São Paulo. Dados obtidos em DATASUS-CNES (2016).

De acordo com as normas médicas, todos os exames médicos baseados em imagens devem ser analisados por um especialista que emitirá um laudo sobre o que pode ser observado nas imagens capturadas (Oliveira et al., 2012). Diariamente no Brasil são realizados milhares de exames e laudos. Somente no primeiro semestre de 2014, no estado de São Paulo, foram realizados 365.203 exames entre radiografias, ressonâncias magnéticas, tomografias computadorizadas e ultrassonografias, o que dá uma média aproximada de 2.028 exames por dia, ou seja, 84 exames por hora (DATASUS-SIAB, 2016). Isto representa um enorme volume de dados textuais e pictóricos que é gerado diariamente. Os dados acumulados nestes laudos proveem uma rica fonte de informação para geração de estatísticas populacionais, para suporte à tomada de decisão médica e apoio na definição de políticas públicas de saúde.

Em 2007, o Conselho Federal de Medicina no Brasil publicou a resolução CFM Nº 1.821/07[1] que aprova as normas técnicas concernentes à digitalização e uso dos sistemas informatizados para a guarda e manuseio dos documentos dos prontuários dos pacientes, autorizando a eliminação do papel e a troca de informação identificada em saúde. Esta decisão considerou os avanços da tecnologia da informação e de telecomunicações, que oferecem novos métodos de armazenamento e transmissão de dados. Leah (2009) destaca que iniciativas semelhantes acorreram nos Estados Unidos em 2004 e que são muito difundidas na Europa, Austrália e em outros lugares. Desta forma, estas iniciativas facilitam o acesso dos laudos para profissionais da área de saúde, pacientes, e também, para os profissionais da área de tecnologia da

[1]http://www.portalmedico.org.br/resolucoes/cfm/2007/1821_2007.htm

informação, eliminando o custoso e demorado processo de digitalização de papéis e contribuindo com processo de recuperação dos dados dos laudos.

Os laudos radiológicos contêm uma fonte rica e prolífica de informações que caracterizam a condição médica do paciente que quando associados às imagens e à clínica permitem o diagnóstico correto. Todavia, uma grande porcentagem das informações contidas nos laudos não está estruturada mas sim em forma de texto livre, o que dificulta o processo de busca, armazenamento e recuperação dessas informações clínicas e inviabiliza a aplicação de métodos clássicos para mineração e análise desta enorme fonte de dados (Rao et al., 2006).

Para amenizar o problema de falta de estrutura das informações, alguns esforços estão sendo empreendidos na tentativa de estruturar as informações dos laudos de uma determinada especialidade durante a sua elaboração, mas sem muito sucesso. Os gestores encontram resistência por parte de muitos profissionais, que consideram o preenchimento de uma interface estruturada mais lento e limitado, interferindo em sua produtividade, comparado ao texto livre no qual o profissional tem liberdade de descrever o que ele observa nas imagens (Heinze et al., 2001a). Algum sucesso tem sido atingido com ferramentas semiestruturadas, que fornecem modelos de parágrafos pré-formatados que são completados pelo profissional com parâmetros digitados ou sugeridos pela própria ferramenta. Porém, o resultado final deste processo ainda é um texto que pode ser alterado pelo profissional. Além disso, os modelos de laudos sofrem uma evolução com o decorrer do tempo, o que significa que em uma mesma base de dados são encontrados laudos com diferentes modelos textuais.

Dentro deste cenário da necessidade de obter informações de laudos médicos, em 2013 foi estabelecido um termo de cooperação técnica entre a Fundação Instituto de Pesquisa e Estudo de Diagnóstico por Imagem (FIDI[2]), a Universidade de São Paulo (USP) e a Universidade Tecnológica Federal do Paraná (UTFPR). A FIDI atualmente é responsável pela realização de 340 mil exames por mês, em mais de 70 unidades espalhadas pelo Brasil. Em 2014 a fundação atingiu a marca de 4 milhões de exames por ano (FIDI, 2016). Este volume de exames gerou uma enorme base de dados de laudos e uma demanda por informações sobre a população que estava sendo examinada. Com o objetivo de extrair essas informações dos laudos foi criado um projeto piloto batizado de Minera®. O capítulo 3 apresenta de forma mais detalhada os resultados das atividades deste projeto. Por agora, vale destacar alguns desafios que foram identificados e que revelam a relevância da atual tese em buscar uma representação de informações não-estruturadas de laudos de exames radiológicos utilizando modelagem interativa de mapas conceituais.

[2]http://fidi.org.br/quem-somos/

1.1 Os desafios da análise de laudos médicos

Além da falta de estrutura, o conteúdo textual dos laudos oferece desafios para sua análise gramatical ao apresentarem erros ortográficos e uma linguagem muito específica e direta. Durante a descrição de um órgão ou patologia, o especialista usa palavras chaves que descrevem a situação observada, sem necessariamente, aplicar pronomes, verbos ou preposições, que trariam uma maior completude linguística ao texto. Nguyen e Patrick (2016) afirmam que a mineração de textos do domínio médico é mais difícil que de outros domínios gerais devido ao grande número de palavras desconhecidas e de dados numéricos. Com isto, um analisador gramatical tem dificuldades para identificar o sujeito, o predicado e outras partes das sentenças contidas nos laudos. Assim, esta incompletude linguística dificulta a aplicação de técnicas de extração de informação utilizando processamento de linguagem natural (PLN) e ontologias (Heinze et al., 2001b).

A Figura 3 mostra um exemplo de laudo de Ultrassonografia Transvaginal. É possível notar alguns problemas com este texto como a sua corretude ortográfica, o tratamento de sinônimos, a falta da completude gramatical, o tratamento de valores numéricos, a identificação de valores multidimensionais, a representação de sentença negativa, o compartilhamento de valores e a relação do conceito com o contexto.

ULTRASSONOGRAFIA TRANSVAGINAL
Bexiga vazia.
Útero visualizado (histerectomia sub-total). O colo mede: 3,1 x 3,0 x 1,8 cm.
Ovário direito: Medindo 3,1 x 2,2 x 2,3 cm nos seus maiores eixos. Volume de 3,4 cm^3.
Aprsentando uma imagem cistica, de aspecto simples, medindo 21 mm (funcional?).
Ovário esquerdo: nao visualizado (grande interposicao gasosa). Ausência de líquido livre na escavação retro uterina. Não evidenciam-se massas ou tumores nas regiões anexiais.
CONCLUSÃO
Cisto em ovario direito.

Figura 3: Exemplo de texto de laudo de Ultrassonografia Transvaginal

As próximas seções apresentam algumas considerações sobre estes desafios e sua contextualização dentro da atual literatura.

1.1.1 Corretude ortográfica

No laudo mostrado anteriormente verifica-se alguns problemas ortográficos que são listados na Tabela 1. Analisando outros laudos, foram encontradas, por exemplo, outras combinações para palavra não: ñao, naõ e ãno. Serapião (et al., 2010) verificou que em um conjunto de 22.247 laudos de mamografia, foram identificados 4.435 termos diferentes, sendo que 934 (21%) estavam com a grafia incorreta.

Tabela 1: Erros ortográficos encontrados em laudos de ultrassonografia transvaginal e sugestões de correção

Erro	Correção
sub-total	subtotal
Aprsentando	Apresentando
cistica	cística
nao	não
interposicao	interposição
ovario	ovário

Estes erros ortográficos fazem com que cada variação incorreta de um termo seja interpretado como um novo termo no conteúdo, ou seja, um novo conceito ou *token*, dependendo da técnica de mineração utilizada. Um dicionário pode ser utilizado para a correção de erros ortográficos. Contudo, os laudos possuem alguns termos muito específicos que podem não ser encontrados em um dicionário de propósito geral.

1.1.2 Tratamento de sinônimos

É possível verificar que em um mesmo laudo há diferentes estilos de escrita. Ao descrever o colo do útero, é utilizada a voz ativa para formar a frase: O colo mede. Enquanto que para descrever o ovário direito é utilizado gerúndio para descrever a mesma característica, sua medida. Tanto o termo 'mede' quando o termo 'medindo' são flexões do verbo medir e podem ser relacionados entre si. No entanto, existem sinônimos mais complexos que não são resolvidos com flexões ou lematização, como por exemplo:

i) "Útero não identificado" == "Útero não caracterizado"

ii) "Ovário direito não visualizado" == "Ovário direito ausente"

iii) "Cisto em ovário direito" == "Ovário direito com formações císticas"

Os dois primeiros exemplos parecem mostrar sinônimos em uma linguagem mais geral e o terceiro exemplo mostra termos específicos do domínio médico. Tomando o primeiro exemplo e analisando as traduções dos termos 'identificado' e 'caracterizado' no banco de dados de palavras WordNet.Pr[3], *identified* e *characterized* não apresentam nenhuma relação semântica. O que leva à conclusão que a utilização destes termos como sinônimos foi definida de forma particular, pelos especialistas da área médica. Logo, esta relação só será identificada a partir de uma ontologia específica que a descrevesse. O mesmo acontece com os sinônimos 'ausente' e 'não visualizados', que não são suportados pela WordNet.Pr. O termo 'não existente' é o sinônimo mais próximo para 'ausente'.

[3] http://wordnetweb.princeton.edu/perl/webwn

1.1.3 Completude gramatical

A linguagem utilizada em laudos é direta e não utiliza, em muitas vezes, as estruturas cultas. Dessa forma, observa-se a omissão de verbos e outros complementos linguísticos que trariam uma maior completude no sentido das sentenças (Nguyen e Patrick, 2016). Algumas abordagens como Ding (2003) e Mashyastha (2003) usam técnicas para identificação de sujeitos, predicados e relações entre objetos diretos e indiretos. Essas técnicas conseguiriam interpretar corretamente uma sentença como:

"O volume do ovário direito é de 3,2 cm³."

Porém, como essas técnicas são dependentes das estruturas linguísticas, teriam dificuldades de identificar as relações em uma sentença como:

"Ovário direito: Medindo 3,1 x 2,2 x 2,3 cm nos seus maiores eixos.
Volume de 3,4 cm³."

Zhou (2003) propõe uma abordagem híbrida que tenta identificar os elementos linguísticos presentes na sentença e estabelecer suas relações, porém, se ela falhar, alguns padrões são utilizados para tentar extrair as informações.

1.1.4 Tratamento de valores numéricos

Além da discriminação de órgãos e patologias, os laudos médicos registram algumas características fisiológicas e patológicas através de valores numéricos. Muitos valores são associados às suas respectivas unidades de medida. Como por exemplo, a pressão arterial é expressa em termos da pressão sistólica sobre a pressão diastólica, e é medida em milímetros de mercúrio (mmHg). O pulso pode ser medido em batimentos por minuto (bpm). Já o peso pode ser registrado em quilograma, em libra, em onça, e etc. A temperatura corporal pode ser medida em graus Celsius (C°) ou graus Fahrenheit (F°). O volume pode estar em milímetros cúbicos (mm^3), em centímetros cúbicos (cm^3) ou em metros cúbicos (m^3), ou ainda, em polegadas cúbicas. Estes são somente alguns exemplos que demonstram que a extração de valores numéricos dependem diretamente da identificação da unidade de medida em que o valor está representado.

Alguns trabalhos como os de Zhou (2006) e de Honorato (2008) assumem que os valores numéricos das características extraídas do texto estarão sempre na mesma unidade de medida. Essa abordagem não seria um problema para valores que realmente tendem a ser expressos usando as unidades de medidas regionais. No entanto, analisando alguns laudos de ultrassonografia, observou-se que características como o volume e a dimensão de órgãos podem estar expressas tanto em mm^3 quanto em cm^3, e que a área de um cisto, por exemplo, pode estar em mm^2 ou cm^2, dependendo muitas vezes do profissional que realizou a anotação ou do montante do valor observado.

1.1.5 Identificação de valores multidimensionais

Algumas características numéricas encontradas nos laudos podem ser unidimensionais, bidimensionais ou tridimensionais, dependendo da técnica utilizada para a aquisição das imagens do exame. Aparentemente, a identificação das dimensões de um conceito pode ser uma tarefa com um certo grau de complexidade. As sentenças a seguir apresentam características tridimensionais e unidimensionais, respectivamente.

"Ovário direito: Medindo 3,1 x 2,2 x 2,3 cm"
"Volume de 3,4 cm³".

Soderland (1999) e Ciravegna (2003) utilizam expressões regulares para a identificação destas ocorrências, E Zhou (et al. 2006) utilizam o resultado dos etiquetadores sintáticos para extrair determinados padrões.

1.1.6 Representação de sentença de negação

A extração de informações estruturadas de um documento não é uma tarefa trivial, como tem sido discutido até aqui. Ainda mais quando os documentos apresentam sentenças de negação. No trecho do laudo a seguir é mostrado um exemplo desse desafio:

" Fundo de saco: Ausência de coleções anormais..."

A sentença acima descreve que numa região denominada *fundo de saco* não foram observadas coleções anormais. A expressão 'coleções anormais' indica um problema, uma situação patológica. O termo 'Ausência' indica que tal patologia não foi observada. Detectar esta negação é importante para que os métodos de recuperação de informação não apresentem alto grau de falsos positivos.

Chapman (et al. 2001) propõe uma abordagem baseada em expressões regulares que identificam a presença de termos de negação em uma sentença e anota a sentença como negativa. É uma abordagem simples, mas que apresenta alto grau de dependência da linguagem e do domínio. O livro de Dowty (1994) reúne uma coleção de artigos de importantes autores da área de análise semântica utilizando recursos mais formais da lógica clássica para tratamento de itens de negação.

1.1.7 Relação do conceito com seu contexto

Ciravegna (2001) discute que a identificação de um conceito dentro de um documento com somente um assunto é muito mais fácil do que em documentos em que são tratados diversos assuntos. Nos laudos médicos são anotados os aspectos de vários órgãos observados durante a realização do exame. Logo, nos laudos são encontrados diversos assuntos. Assim, ao identificar o conceito 'volume', por exemplo, é necessário identificar qual órgão ou patologia ele referencia. No trecho do laudo a seguir é possível verificar a ocorrência do conceito 'medindo' duas vezes.

... Ovário direito: Medindo 3,1 x 2,2 x 2,3 cm nos seus maiores eixos. Volume de 3,4 cm³. Aprsentando uma imagem cistica, de aspecto simples, medindo 21 mm...

Trabalhos como de (Autor1, 2010) e (Autor2, 2011) utilizam expressões regulares para detectar determinados padrões de ocorrência de termos e estabelecer a ligação entre os conceitos. Os padrões devem ser previamente identificados e adicionados ao processo de extração de informações. Outros autores utilizam ontologia que determinam as relações entre os conceitos de um determinado domínio.

1.2 Motivação

Analisando os desafios discutidos anteriormente, observa-se que as soluções indicadas exigem um alto grau de conhecimento especializado em processamento de linguagem natural e também no domínio do texto analisado. Muitas abordagens exigem grande esforço da equipe para a montagem do arcabouço conceitual necessário nos procedimentos de análise do texto e extração das informações. Dada a complexidade do processo e destes recursos conceituais, muitas empresas são desencorajadas por não encontrar os profissionais capacitados para a elaboração do projeto ou por não possuírem recursos financeiros para contratar profissionais com tal grau de especialidade.

A ideia central deste trabalho surgiu da experiência que seu autor teve no desenvolvimento do projeto Minera®. Neste projeto foram criados um dicionário, uma ontologia e um analisador de textos para extração dos dados conforme o dicionário e a ontologia definidos. Os resultados foram satisfatórios e o projeto foi usado para alguns levantamentos estatísticos sobre a população que realizou ultrassonografias transvaginais em determinado período, como descrito na seção 3.7, na página 31.

Ao tentar aplicar o projeto Miner@ em outras modalidades de exames radiológicos como tomografia e ressonância magnética e outras especialidades como neurologia ou ortopedia, percebeu-se que todo o trabalho de análise de domínio e definição da ontologia deveria ser repetido para cada modalidade e especialidade, conforme observa . Logo, os custos de implantação seriam compatíveis com os custos das primeiras fases do projeto, e pouco se aproveitaria.

Tendo em vista que os laudos usam uma linguagem descritiva, direta e sem muito formalismo gramatical, como será mostrado no Capítulo 4, surge a pergunta:

É possível extrair informações estruturadas dos laudos sem usar este arcabouço de recursos para o processamento de linguagem natural (PLN)?

Todas as atividades desta tese foram desenvolvidas para demonstrar que a resposta a esta pergunta é afirmativa. Sim, é possível desenvolver sistemas de extração de informação e pesquisa estruturada em laudos com baixo custo de acoplamento e implantação. Com o uso de tal abordagem, grandes repositórios de laudos médicos podem ser analisados e utilizados para a gerar conhecimentos sobre a população examinada e também sobre patologias e fisiologias populacionais.

Com o desenvolvimento desta pesquisa pretende-se facilitar a extração de informações de laudos e a criação de uma rica fonte de conhecimento, possibilitando assim, traçar o perfil dos usuários dos serviços de diagnósticos e quantificar as doenças mais relevantes, permite a criação de índices que podem auxiliar na indicação de prioridades em saúde e auxiliar no estabelecimento de políticas públicas de saúde regionais e nacionais. Os dados para este propósito estão nos enormes volumes de laudos armazenados atualmente em instituições públicas e privadas, que mantém as bases com o propósito de consulta e para amparo legal.

1.3 Objetivos

O principal objetivo deste trabalho é apresentar uma abordagem de extração de informações não estruturadas de laudos médicos utilizando mapas conceituais construídos pelos próprios especialistas, a partir de laudos já existentes. A Figura 4 mostra um exemplo de mapa conceitual montado a partir da ligação dos termos presentes nas três primeiras linhas do laudo mostrado na Figura 3.

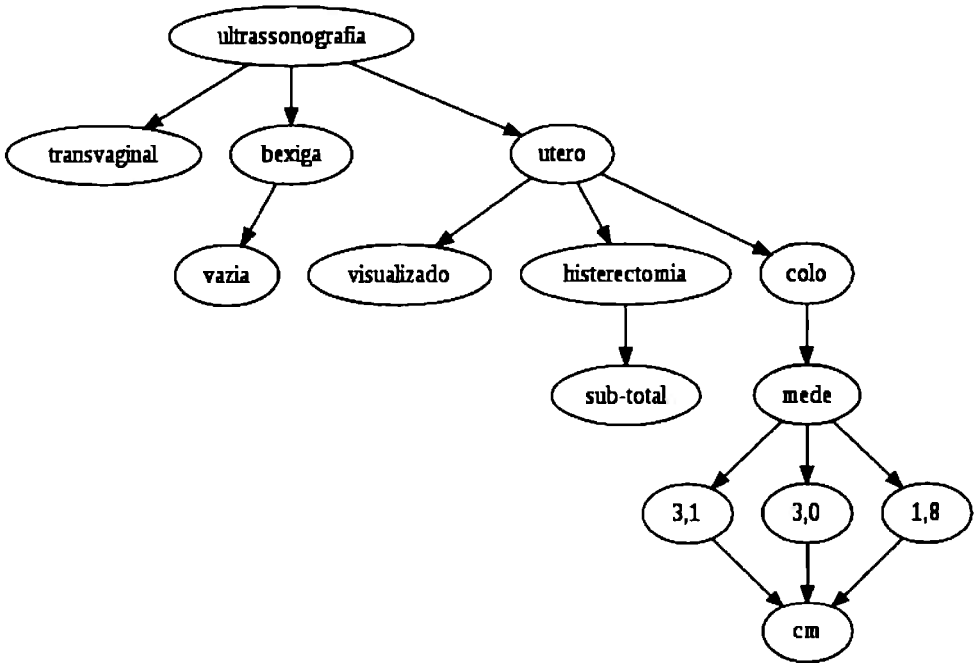

Figura 4: Exemplo de um mapa conceitual a partir dos termos presentes no laudo

Embora exista um arcabouço de outras técnicas disponíveis na literatura para a extração de informações em textos, como pode ser verificado no livro de Aggarwal e Zhai (2012), o desafio dessa proposta é desenvolver um método prático de modelagem e extração de informações, que não exija conhecimento especializado sobre PLN ou ontologia por parte do operador. As ligações entre as palavras deverão ser realizadas pelo especialista médico de uma forma visual e intuitiva. E o método se encarregará de extrair o máximo de informações sobre cada interação, analisando os termos ligados, e os termos precedentes, intermediários e subsequentes a cada ligação.

De uma forma mais abrangente, esta proposta almeja facilitar o processo de extração e estruturação das informações contidas no grande volume de dados textuais disponíveis em muitas instituições médicas brasileiras e que, atualmente, são subutilizados. Isto pode ser alcançado reduzindo a complexidade da implementação e implantação de processos de extração de informação, o que não é observado em outras metodologias.

Os objetivos específicos deste trabalho são:

- Definição de um modelo de mapa conceitual para o armazenamento dos conceitos e suas relações nos laudos, de forma interativa. Este modelo é o resultado da análise e da união de todos os mapas criados pelos especialistas. Ele armazena os conceitos e as características locais de cada relação no momento de sua definição pelo especialista.
- Desenvolvimento do interpretador do modelo que será capaz de estabelecer as mesmas relações em outros laudos dentro da mesma especialidade.
- Desenvolvimento das ferramentas de apoio: i) Um ambiente para a construção dos mapas pelos especialistas; ii) Um ambiente para gerenciamento dos mapas construídos; iii) Uma ferramenta para seleção dos mapas e construção do modelo conceitual; iv) Uma ferramenta para teste e validação do modelo conceitual.

1.4 Metodologia

A atual pesquisa iniciou-se com o desenvolvimento do projeto Minera® e aplicação de técnicas clássicas de PLN, usando dicionários, tesauros e ontologias. Devido ao alto custo de replicação do projeto para outras modalidades e especialidades de exames, os pesquisadores iniciaram uma busca por alternativas que envolvesse um menor número de profissionais especialistas no projeto.

O primeiro passo foi fazer uma revisão bibliográfica e analisar as atuais abordagens para representação e extração de informações não estruturadas, como apresentada no Capítulo 2.

Dada a particularidade dos textos dos laudos, formulou-se então uma solução usando mapas conceituais, que podem ser gerados pelos próprios especialistas.

Utilizando textos de laudos já escritos, foram realizadas ligações entre os termos de forma a representar suas relações, como por exemplo, ligando volume com útero, criando assim, um mapa conceitual do laudo. Vários dados sobre estas ligações foram armazenados.

Um modelo conceitual foi gerado unindo vários mapas conceituais e os dados coletados durante as ligações. Este modelo foi então usado para determinar a probabilidade de possíveis ligações entre conceitos presentes em laudos semelhantes.

Para a validação do modelo foram analisadas sua precisão e sua cobertura dado um conjunto de controle composto por laudos anotados.

1.4.1 Classificação da pesquisa

Tendo como base as discussões sobre metodologia de pesquisa de Silva (et al., 2001), a atual dissertação é classificada da seguinte forma:

De natureza aplicada por gerar conhecimento para aplicação prática e dirigidos à solução do problema de recuperação de informações não estruturados em laudos médicos.

Quantitativa por apresentar uma análise morfológica textual dos laudos e o índice de precisão e acurácia do modelo aplicado ao conjunto de teste.

Quanto aos objetivos, a atual pesquisa é exploratória, pois levanta uma hipótese para construir um modelo de representação de informações não estruturadas de laudos médicos usando mapas conceituais e explora como esta técnica pode ser aplicada.

Quanto aos procedimentos técnicos, nesta tese foram utilizados o levamento bibliográfico, o estudo do caso Minera® a realização de experimentos para a fundamentação dos resultados do Analaudos.

1.5 Contribuições

- um novo método de extração de informações a partir de fontes não estruturadas, com baixo acoplamento de tecnologias e fontes externas de conhecimento;
- uma modelagem visual e intuitiva para anotação de conceitos em conteúdos textuais, gerando um modelo em grafo conceitual, sem a necessidade de escrita de expressões de extração;
- permitir a seleção de laudos que satisfaçam determinadas características nominais, discretas ou contínuas expressas em seu conteúdo, conforme o mapeamento conceitual realizado pelo modelo;
- contribuir com alguns pontos carentes da literatura, como a análise automatizada de textos que contenham dados numéricos, multidimensionais e em diferentes escalas;
- possibilitar a análise de similaridade conceitual dos laudos, aplicando técnicas de análise de similaridade entre mapas.

1.6 Organização do trabalho

O próximo capítulo apresenta uma análise do estado da arte em extração de informações e seu relacionamento com os objetivos dessa proposta. O capítulo 3 apresenta uma descrição detalhada das atividades e resultados do projeto Minera®. O capítulo 4 apresenta um estudo sobre a morfologia textual de laudos de ultrassonografia transvaginal, indicando as características de um texto no qual o modelo desta tese seja aplicável. O capítulo 5 apresenta o desenvolvimento da proposta desta tese, o Analaudos.

2 Fundamentação

A Economia do Conhecimento e a Sociedade da Informação aumentam a demanda pelo acesso e processamento de dados gerados por diferentes fontes (Brasil, 2002). No entanto, a integração dessas fontes não é uma tarefa trivial. Goh (1996) destaca em sua tese os conflitos sintáticos e semânticos entre os dados provindos de bases heterogêneas ou distribuídas. As soluções para estes conflitos são implementadas dentro dos sistemas que integram as diversas fontes. Esses sistemas possuem as informações sobre os diferentes esquemas de dados e os algoritmos para conversão e adaptação dos mesmos.

Com o desenvolvimento tecnológico e principalmente dos sistemas de informação, torna-se cada vez mais acessível e comum editar e manter os documentos em sua forma digital. Esses documentos são armazenados em diferentes formatos. E a integração das informações contidas nesses documentos não estruturados é uma tarefa ainda mais difícil, pois os conflitos semânticos são maiores e dependentes do domínio da informação e da linguagem utilizada.A área de extração de informações é responsável pelo estudo e desenvolvimento de técnicas capazes de extrair informações contidas em fontes textuais. Os estudos nesta área estão intrinsecamente ligados ao estudo da linguística que buscam extrair as semânticas expressas em sentenças construídas a partir de uma linguagem específica. Na Ciência da Computação, o Processamento de Linguagem Natural (PLN) é a área responsável pela extração e representação da semântica utilizando estruturas capazes de serem interpretadas por computador. Uma parceria bem sucedida foi estabelecida entre o PLN e a ontologia, uma vez que, na Ciência da Computação, a ontologia é a área responsável pela representação, compartilhamento e reutilização do conhecimento (Freitas e Vieira, 2008). Tanto os recursos do PLN como da ontologia são bastante dependentes do domínio e da linguagem utilizados.

A linguagem escrita é um conjunto de símbolos que, quando agrupados seguindo regras específicas, é capaz de registrar informações, pensamentos, entre outros elementos da comunicação. A análise do significado dos componentes textuais auxilia na busca pela semântica dos conteúdos. Ullmann (1962) registra que no século I a.C. Aristóteles já apresentava as primeiras iniciativas para classificar metáforas. Atualmente, Feldman, R. e Sanger, J. (2006) separam o processo de extração de informações e de conhecimento de documentos textuais em quatro fases bem definidas, mostradas na Figura 5.

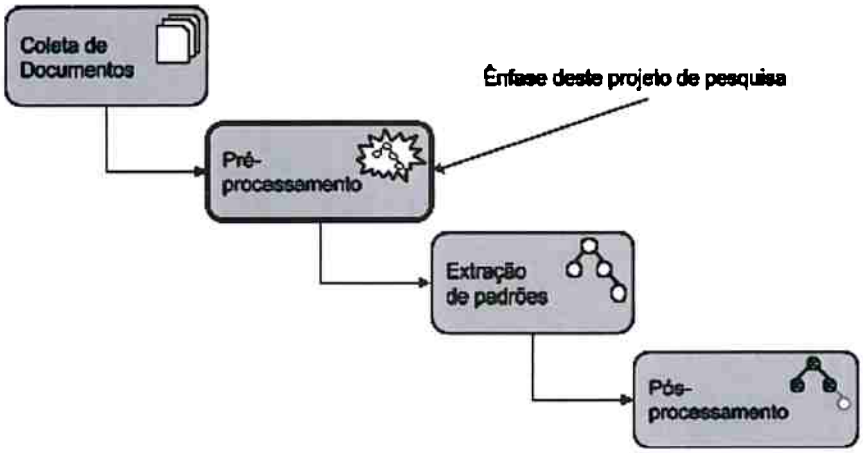

Figura 5: Processo de extração de informações de documentos textuais, segundo Feldman, R. & Sanger, J. (2006)

A fase de coleta de documentos compreende as atividades de acessar os documentos em suas diferentes fontes e selecionar os documentos que serão utilizados nas próximas fases do processo de extração de informações. A fase de préprocessamento é a fase mais importante do processo, e também a mais dependente do domínio. Os objetivos do atual projeto se concentram nesta fase. Durante o pré-processamento ocorrem as tentativas de estruturação das informações, nas quais o conteúdo é analisado para identificar as relações entre suas partes e inferir algumas estruturas. A fase de extração de padrões é realizada sobre as informações já extraídas e estruturadas, e conta com diversas técnicas bastante consolidadas para classificar e agrupar o conteúdo, como as bayesianas, k-means, entre outras (Alpaydin, 2010). A fase de pós-processamento é responsável pela análise e validação dos padrões identificados.

2.1 Pré-processamento para estruturação de dados

Lapa (2013) apresenta uma evolução histórica das técnicas aplicadas para o processamento de conteúdos textuais. O desenvolvimento do hardware foi um fator determinante para os avanços nesta área que inicialmente recebeu contribuições de métodos estatísticos e probabilísticos, anteriores aos anos 80, e posteriormente, de métodos linguísticos, nos anos 80, e de aprendizado de máquina, nos anos 90. A técnica mais comum para a extração de informações a partir de conteúdo não estruturado baseia-se em características que podem ser abstraídas a partir de um texto, uma imagem ou mesmo, de um som. Essas características são obtidas a partir de algoritmos específicos que analisam determinados aspectos do conteúdo e converte-os em valores nominais ou numéricos. A representação do conteúdo em uma forma linear ou discreta possibilita, entre outras análises, a análise de similaridade entre conteúdos. Uma vez extraídas as características, as técnicas de identificação de padrões podem ser aplicadas, independentemente da natureza do conteúdo. Contudo, a análise e interpretação dos padrões na fase de pós-processamento deverá levar em consideração sua natureza.

Zhai (20??) discute sobre duas grandes categorias de técnicas para extração de informações textuais, a simbólica utilizando lógica e regras de inferência e as abordagens estatísticas.

2.1.1 As características de um texto

Em um texto, cada palavra pode ser analisada como uma característica, e uma classe de texto pode ser representada por um conjunto de palavras chaves presentes em seu conteúdo. Oleynik (2013) desenvolveu um classificador bayesiano para classificar laudos de anatomia patológica capaz de inferir a topografia e a morfologia de um câncer na Classificação Internacional de Doenças para Oncologia (CID-O). Neste trabalho, um conjunto de narrativas clínicas é manualmente classificado e serve para o treinamento do classificador. Cada classe é determinada pela presença de um conjunto de palavras chaves que identificam as categorias de um câncer. Como prática importante em trabalhos que envolvem PLN, algumas transformações são realizadas antes do processamento do conteúdo, levando em consideração sua morfologia, lexicografia e sintaxe.

2.1.2 Análise léxica

A análise morfológica é responsável por determinar os itens que compõe o conteúdo, ou seja, as palavras, pontuações e sentenças. Isto permite a separação do conteúdo em unidades léxicas (*tokens*) que serão analisadas posteriormente de forma individual ou conjunta. Muitos o verão com outros olhos.

As transformações léxicas são importantes para a normalização do conteúdo textual, tornando-o mais limpo e com menos ruídos causados por variações da linguagem. Destaca-se, com relação à atual proposta, as técnicas de derivação (*stemming*) e de lematização (*lemmatization*). Ambas técnicas buscam reduzir as palavras para um radical invariante. A derivação utiliza processos heurísticos que removem prefixos e sufixos de uma determinada palavra, resultando em um lexema não reconhecido pelo dicionário da linguagem, mas que mantém, de certa forma, a semântica do termo. A lematização busca representar uma palavra em sua forma canônica: o lema, ou seja, um radical que pertence ao dicionário e que representa uma classe de possíveis variações nominais e verbais. Oleynik (2010) destaca que muitas técnicas de PLN têm seus desempenhos prejudicados ao serem aplicadas em documentos médicos por causa da especificidade da linguagem, uma vez que muitas ferramentas são treinadas em textos jornalísticos ou de propósito geral.

Neto (2011) em seu trabalho com três bases de 5000 laudos cada uma, usa somente a análise das frequências dos termos e de n-gramas. Ele identifica os termos singulares mais frequentes e depois efetua uma busca por estes termos na lista de n-gramas mais frequentes. Ao identificar combinações de termos mais relevantes, ele infere um relacionamento semântico, como por exemplo, os termos 'joelho ligamento posterior' e 'joelho ligamento lateral' dão origem aos conceitos 'joelho', 'ligamento', 'lateral' e 'posterior'.

Serapião (et al. 2010) realiza dois experimentos com 22.247 laudos de mamografia para avaliar sua qualidade informacional usando somente as informações léxicas. Nesses experimentos, é verificada a

variação léxica das seções de descrição e conclusão dos laudos e analisada a corretude ortográfica de cada léxico.

Também usando somente as informações léxicas de laudos, Barbosa (2013) analisa a densidade e a aglutinação léxica dos laudos de três instituições diferentes para identificar termos compartilhados e propor uma superestrutura padrão para os laudos.

2.1.3 A análise sintática

Ao analisar somente a presença de palavras chaves em um texto, são desprezados a localização e o relacionamento destas palavras com o restante do conteúdo. Uma análise sintática contribui para o entendimento dos componentes e do seus relacionamentos, contribuindo, posteriormente, para a extração da semântica do conteúdo. Para isto, várias características são extraídas de um conteúdo textual. Geralmente, estas características são representadas na forma de etiquetas que são definidas para um termo ou para um determinado conjunto de termos. Destacam-se três tipos de etiquetas comumente utilizadas nos trabalhos de PLN:

- as **etiquetas léxicas** são atribuídas a partir das classes gramaticais das palavras, como substantivo, verbo, artigo, adjetivo, preposição, pronome, advérbio, conjunção, interjeição e suas subcategorias. O termo POS-tagger (do inglês *part-of-speech tagger* ou etiquetador das partes do discurso) é atribuído ao gerador dessas etiquetas, que geralmente é treinado com textos de propósito geral. Esses etiquetadores podem usar diversas técnicas para gerar suas etiquetas, dentre elas, as técnicas estocásticas que determinam a probabilidade de um termo ou uma classe gramatical preceder ou suceder outro; as técnicas baseadas em regras gramaticais da linguagem que determinam a classificação de uma certa sequência de termos; e as técnicas baseadas em dicionários. É importante destacar que para o PLN, um dicionário é um conjunto de termos que são reconhecidos por uma linguagem.

- as **etiquetas sintáticas** são atribuídas a partir de combinações de estruturas gramaticais mais complexas, como sujeito, predicado, frases verbais e frases nominais. Estes etiquetadores são conhecidos como chunks, que em português seria traduzido como pedaços, uma referência ao seu trabalho de agrupar partes de uma sentença que juntas expressam algum significado. Este agrupamento pode ser realizado utilizando as etiquetas léxicas e aplicando regras gramaticais para identificar certos padrões de sequências da linguagem.

- as **etiquetas semânticas** referem-se ao significado dos termos, das sentenças e sua relação dentro de um determinado domínio. Os etiquetadores semânticos geralmente utilizam uma ontologia para associar um termo ou um conjunto de termos ou conceitos, ou utilizam a hierarquia de um tesauros para identificar a generalização ou a especialização de conceitos.

O trabalho de Haythornthwaite (2007) é um exemplo do uso de técnicas simples para identificação da semântica de conteúdos textuais. Ele analisa as ocorrências de características léxicas e sintáticas para identificar os assuntos das conversas registradas em um fórum de discussão de um ambiente educacional. Todas as mensagens são analisadas previamente. Cada termo da mensagem recebe sua etiqueta léxica. Uma regra sintática personalizada é utilizada para extrair frases nominais. Somente os adjetivos e substantivos são analisados. Os verbos, artigos e outros elementos são desprezados. O número de ocorrências de cada substantivo ou frase nominal é analisado e uma lista é construída, iniciando pelo item mais frequente. Assim, segundo o autor, a análise dessa lista permite a identificação dos assuntos mais discutidos e de outros aspectos da interação entre os membros do fórum. Os significados dos termos e suas relações não são analisados, caracterizando-se, assim, um método basicamente estatístico.

Zhou (2005) também usa uma técnica sintática bastante simples para otimizar o número de busca por conceitos em registros médicos de um paciente. Ele observa que muitos termos médicos são compostos por várias palavras. Assim, ele identifica candidatos para termos compostos aplicando as etiquetas léxicas em cada termo do conteúdo e depois buscando por agrupamentos específicos de adjetivos e substantivos. Desta forma, evita-se a busca por todas as combinações de n-gramas para encontrar um termo médico.

Ijntema (et al. 2012) compara a eficácia entre análises sintáticas mais complexas utilizam padrões léxico-sintáticos e léxico-semânticos. Os padrões são especificados a partir de expressões que utilizam as etiquetas léxicas, sintáticas e semânticas, previamente agregadas ao conteúdo. Nesses sistemas, uma informação é extraída a partir de expressões que identificam determinados arranjos de termos ou de etiquetas. Como por exemplo, no sistema WHISK (Soderland, 1999) a expressão da Figura 6 extrai de textos de anúncios o valor do aluguel e quantidade de quartos no imóvel. As etiquetas (Digit) e (Number) deverão ser devidamente geradas pelos etiquetadores antes da aplicação do padrão sobre o texto analisado.

Pattern: * (Digit) ' BR' * '$' (Number)
Output: Rental {Bedroom $1} {Price $2}

Figura 6: Expressão de extração no sistema Whisk.

Vários outros autores definem linguagens para construções de expressões para extração de informações como GE NLToolset (Jacobs, 1991), LSPE (Hearst, 1998), CAFETIERE (Black, 2005) e HERMES (Ijntema et al. 2012). As principais diferenças entre as propostas são o nível semântico das expressões, o arcabouço de tecnologia utilizado e o grau de dificuldade na composição das expressões. Uma ou mais expressões podem ser definidas para a extração de um mesmo dado ou conceito, permitindo que algumas variações do conteúdo sejam corretamente interpretadas. Contudo, o especialista do domínio deve ter em mente o formato exato do conteúdo durante a composição das expressões.

Ciravegna (2001) propõe um método de extração de informações a partir de um conjunto de documentos anotados com delimitadores de conceitos (<LUGAR></LUGAR>, <PESSOA></PESSOAL>, etc). O método analisa os termos circunvizinhos a cada anotação e define uma expressão padrão para inserção de cada parte do delimitador, a inicial e a final. Aplicando as etiquetas a cada termo da regra padrão, outras regras mais genéricas são derivadas e testadas no restante do corpus anotado para verificar sua precisão. As regras com melhores resultados são preservadas e utilizadas para extrair os conceitos a partir de outros documentos similares. Dessa forma, o método tenta aprender as expressões capazes de extrair os conceitos delimitados, evitando que o usuário tenha que escrevê-las.

2.1.4 A análise semântica

Para a extração da informação é necessário entender o significado das partes do conteúdo textual para posteriormente analisar seus relacionamentos e inferir o significado do conteúdo por completo. As etiquetas léxicas e sintáticas fornecem os significados linguísticos dos termos, levando em consideração o dicionário e a gramática de uma linguagem. As etiquetas semânticas são responsáveis por proporcionar um entendimento sobre o assunto; por isso, são altamente dependentes do domínio e necessitam de fontes específicas que forneçam sinônimos e conceitos para seu processamento. Breitman (et al. 2007) destacam que as principais fontes de dados semânticos utilizadas nos processos de PLN são os tesauros e as ontologias, que serão discutidos na seção seguinte.

A análise semântica é realizada aplicando as restrições existentes entre os conceitos destacados no conteúdo. Bräscher (2002) exemplifica a análise semântica do conceito 'ação' que pode se tratar de um título de crédito, praticar ação, entre outros sentidos. O que determinará o significado da ação serão os conceitos circunvizinhos que poderão denotar um ato de vender ou um comportamento.

O BioPatentMiner (Mukherjea et al. 2004) é um exemplo de sistema que usa a análise semântica para facilitar a busca de informações sobre patentes médicas. Um conjunto de patentes anotadas e alguns recursos de dicionários médicos formam a sua base semântica. O sistema identifica e classifica os termos relacionados à medicina e integra a esses termos seus respectivos conceitos (classes, subclasses e sinônimos). Dentro de uma patente ele também identifica as relações de inventor, de sucessor e de referência a outra patente. A partir dessa extração o sistema utiliza um motor de inferência para buscar por patentes utilizando termos mais genéricos ou mais específicos, e também utilizando as informações sobre as relações.

Montes-y-Gómez (2000) utiliza a estrutura de grafos conceituais propostos por Sowa (1984) para modelar as informações de um conteúdo textual e realizar buscas semânticas. Seu método consiste em analisar a similaridade conceitual e relacional entre os grafos dos documentos e o grafo de uma frase de busca. Ele não detalha como ocorre o processo de montagem dos grafos, mas mostra a eficiência do método em selecionar documentos com maior relevância semântica, uma vez que a relação entre os conceitos é analisada.

2.2 Representação e armazenamento do conhecimento

Sowa (1984) destaca que no século XIX houve grandes avanços na área da lógica com os trabalhos de George Boole, com sua lógica diádica de valores 0 e 1 e os operadores de disjunção, conjunção e negação. Mais tarde, no mesmo século, Charles Sandres Peirce apresentou suas contribuições para a lógica propondo os operadores de implicação e os quantificadores existencial e universal. Estes foram os fundamentos para a representação de conhecimento que posteriormente foram utilizados na computação. Sowa propõe a representação do conhecimento utilizando grafos conceituais (GC) que são compostos basicamente por conceitos e relações. Para compor sua técnica, Sowa baseou-se nos gráficos existenciais de Peirce, nos gráficos de dependência de Tesnière e nas redes semânticas da Inteligência Artificial.

Outra forma de armazenar conhecimento são os mapas conceituais, que segundo Moreira e Buchweitz (1987) são diagramas bidimensionais capazes de representar conceitos de determinado domínio organizados de forma hierárquica. A hierarquia é determinada baseando-se na própria estrutura da fonte de conhecimento utilizada pelo observador e construtor do mapa.

Os grafos conceituais são facilmente traduzidos para RDF[4] (*Resource Description Framework*), uma linguagem para modelagem e descrição de conceitos. Isto permite a utilização de um arcabouço de tecnologia para manipulação de bases de conhecimento.

O projeto WordNet.Pr (Fellbaum, 1998), desenvolvido pela Universidade de Princeton, nos Estados Unidos, é pioneiro na construção de uma estrutura de sinônimos semanticamente interligados. Trata-se de um grande dicionário eletrônico, em inglês, com aproximadamente 155.000 itens divididos em quatro grandes grupos: substantivos, verbos, adjetivos e advérbios. Cada item é associado aos seus conjuntos de sinônimos. Cada conjunto contém uma descrição de seu significado, uma sentença de exemplo de uso, seu grupo maior e seu subgrupo. Os subgrupos são divididos de acordo com uma semântica simples, como por exemplo, planta, comida, verbo de ação, e outros 47 subgrupos. Um conjunto pode se relacionar com outro através de relações semânticas pré-definidas, permitindo identificar, por exemplo, conjuntos antônimos, entre outras relações como 'é uma parte de', 'é o todo de', 'é um super tipo de' e 'é um subtipo de'. A Figura 7 mostra o resultado na ferramenta online[5] da pesquisa pelas palavras car door e suas relações semânticas com outros conjuntos. O WordNet é bastante útil para projetos que envolvam conteúdos textuais em Inglês e de propósito geral.

[4] http://www.w3.org/TR/REC-rdf-syntax/
[5] http://wordnetweb.princeton.edu/perl/webwn

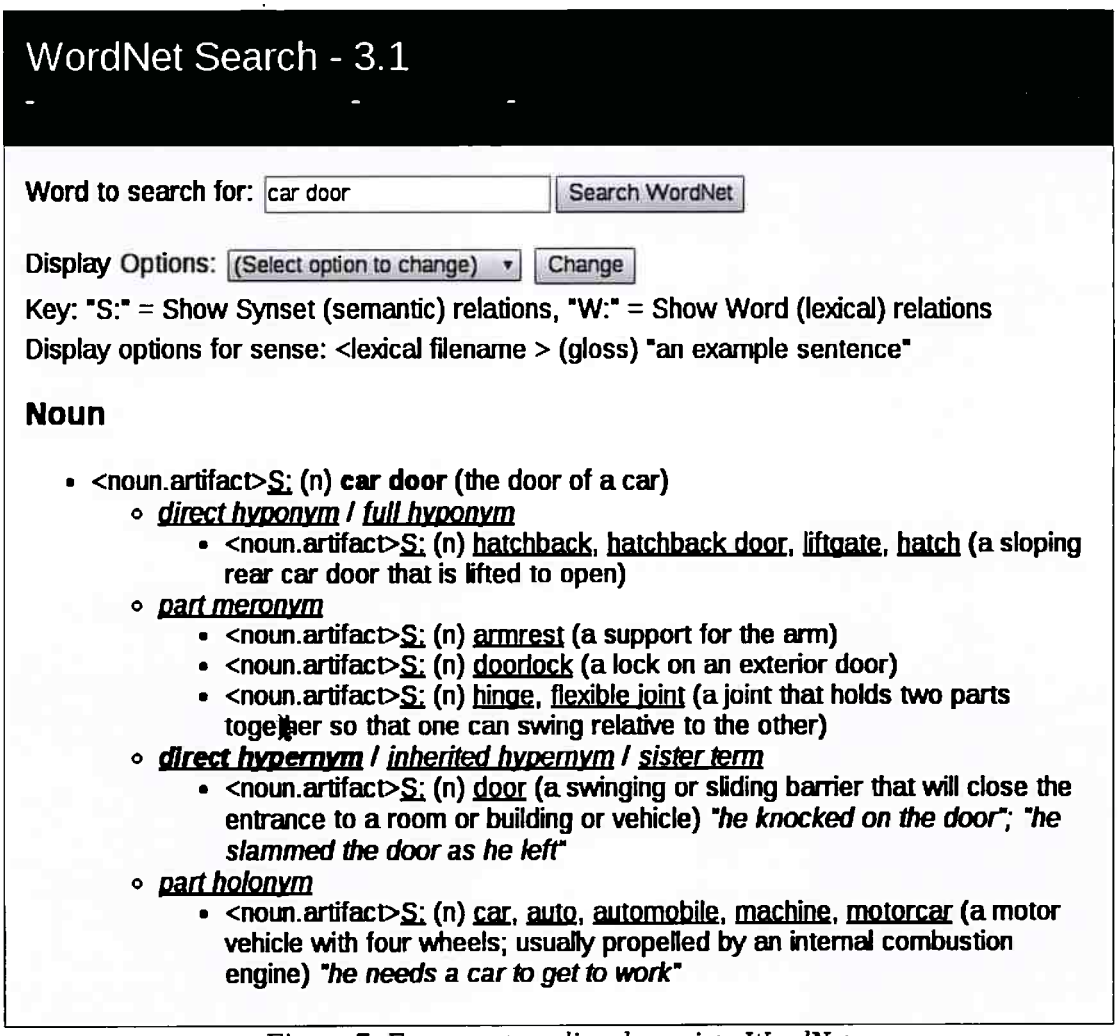

Figura 7: Ferramenta online do projeto WordNet

O projeto WordNet.br (Dias-da-Silva, 2003), desenvolvido por diversas universidades brasileiras parceiras, é uma iniciativa alinhada com a WordNet.Pr, que conta com 5860 verbos agrupados em 3713 conjuntos de sinônimos.

Os dois projetos citados anteriormente são para uso de propósito geral, pois não possuem registros de termos específicos, como por exemplo, da área médica. O projeto UMLS, do inglês, Unified Medical Language System (Bodenreider, 2004), é mantido pela U.S. National Library of Medicine, reúne diversos recursos linguísticos e semânticos sobre saúde, termos médicos e nomes de remédios.

O projeto DeCS, Descritores em Ciências da Saúde (Pellizzon, 2004), mantido pela BIREME[6], foi criado em 1986 a partir da tradução e adaptação do Medical Subject Headings – MeSH[7] (Lipscomb, 2000) produzido pela U.S. National Library of Medicine. O MeSH existe desde 1960, para uso na indexação de documentos tais como: artigos de revistas científicas, livros, anais de congressos, relatórios técnicos, e outros tipos de materiais.

[6] http://www.bireme.br

[7] http://www.nlm.nih.gov/mesh/meshhome.html

As ontologias representam o conhecimento através da definição de classes, relações semânticas e instâncias, e são muito úteis para inferências. Freitas e Vieira (2008) afirmam que as ontologias representam o ponto mais elevado já atingido em termos de representação, compartilhamento e reutilização do conhecimento. Um arcabouço de ferramentas já foram desenvolvidas para a criação, manipulação e consulta de ontologias. A Web Ontology Language (OWL) é uma linguagem formal para a descrição de ontologias (Bechhofer et al., 2004).

Os tesauros são considerados ontologias hierárquicas mais limitadas, com propósitos mais terminológicos, pois possuem uma estrutura mais rígida e categorias de relações específicas. Nele os conceitos relacionam-se com seus significados, sinônimos, antônimos e outros conceitos mais gerais ou mais específicos. Ambos métodos permitem que um determinado termo seja localizado em sua hierarquia e, assim, suas relações são recuperadas, podendo ser um processo interativo e em vários níveis (Breutman et al., 2007). Maynard (2008) destaca que um item pesquisado e não encontrado num tesauro pode ser interpretado como inexistente. No entanto, os tesauros são incompletos e bastante fechados em um determinado domínio.

2.3 Ferramentas de apoio

As atividades de PLN para a realização das análises léxica, sintática e semântica são bem consolidadas na literatura, variando, muitas vezes, na linguagem, no corpus utilizado para o treinamento e análise, e também nos métodos utilizados pelos etiquetadores. Com o propósito de auxiliar um especialista em PLN na montagem do seu ambiente de extração de informações, várias ferramentas foram desenvolvidas.

A ferramenta GATE (Cunningham, 2002), desenvolvida na linguagem Java[8], provê uma infraestrutura modular de propósito geral para a construção de sistemas de extração de informações. Faz parte dos seus módulos diversos etiquetadores léxicos, sintáticos e semânticos, que podem ser configurados e estendidos. Oferece também um módulo para especificação da sequência de execução das atividades, incluindo suporte para manipulação de corpus de testes e de análise. Essa ferramenta é usada por vários outros trabalhos de propósito mais específico e domínio mais restrito, como Circunavega (et al. 2003), Popov (et al. 2004), Zhou (2006), Carvalheira (2007).

2.4 Abordagens para estruturação dos dados de laudos médicos

Duas técnicas se destacam nos trabalhos publicados sobre a extração de dados não estruturados em laudos médicos. As análises estatísticas analisam a ocorrência dos termos dentro dos laudos para tentar classificá-los, já as análises semânticas exploram a relação entre os termos dentro de um mesmo laudo.

[8]http://www.java.com

3 Projeto Miner@

Em 2013 foi estabelecida uma parceria entre as universidades USP e UTFPR e a fundação FIDI (Fundação Instituto de Pesquisa e Estudos de Diagnóstico por imagem) para o desenvolvimento de uma solução para extração de informações dos laudos médicos. A FIDI é a maior prestadora de serviço de diagnóstico por imagem para o SUS. Ela realiza cerca de 340 mil exames ao mês, em 75 unidades SUS em todo Estado de São Paulo. Isto representa um grande volume de laudos que são armazenados com informações importantes sobre a população examinada, mas que, até o momento, vinha sendo subutilizada.

Para viabilizar o desenvolvimento do projeto, o escopo inicial incluiu somente os laudos de ultrassonografia pélvica e transvaginal. Esta modalidade de exame foi escolhida devido a dois fatores principais: a disponibilidade de especialistas para participação no projeto, e a simplicidade das estruturas fisiológicas observadas, ou seja, útero, ovários e anexos. O projeto foi registrado no comitê de ética sob o número 169/11 CEP/ICS/UNIP.

A seguir são descritas as fases do processo desenvolvido para a extração de informações não estruturadas dos laudos de ultrassonografia pélvica e transvaginal.

3.1 Levantamento de algumas perguntas para submissão na ferramenta de busca

Em reunião com o Dr. Harley de Nicola, membro do Comitê Científico da FIDI, foram definidas algumas possíveis buscas que nortearam o desenvolvimento do projeto:

- "Selecionar laudos de pacientes que possuem o ovário menor que 5 cm^3"
- "Selecionar laudos de pacientes que possuem o endométrio com espessura superior a 10 mm"
- "Selecionar laudos de pacientes que possuem mioma"
- "Selecionar laudos de pacientes que possuem mioma submucoso"
- "Selecionar laudos de pacientes que possuem mioma intramural"
- "Selecionar laudos de pacientes que possuem mioma subseroso"
- "Selecionar laudos de pacientes cujo ovário não foi visualizado ou não foi identificado"
- "Selecionar laudos de pacientes que possuem ovário com volume entre 10,1 e 50 cm^3"
- "Selecionar laudos de pacientes com cisto entre 3,1 e 100 cm^3"
- "Selecionar laudos de pacientes com cisto entre 0 e 3 cm^3"
- "Selecionar laudos de pacientes com endometriose"

3.2 Análise e extração do conteúdo dos laudos e fontes de informações sobre anatomia, exames e patologias

Os laudos dos exames são armazenados em um banco de dados relacional. O esquema do banco foi analisado para identificar as tabelas e colunas que seriam utilizadas pela ferramenta de extração.

O conteúdo dos laudos é armazenado no formato RTF.

Foi desenvolvido um módulo de extração do conteúdo dos laudos capaz de conectar-se com o banco de dados e realizar as consultas SQL necessárias. Além disso, o conteúdo RTF dos laudos é processado para a remoção de metadados de formatação e somente o texto plano, sem formatação, é analisado.

3.3 Análise da indexação convencional

Para entender melhor o conteúdo dos laudos e obter parâmetros de referência para comparar os resultados finais do projeto, foi decidido realizar uma indexação convencional. Isto permitiu mostrar a todos do grupo as limitações desses índices e analisar o desempenho de buscas por palavras-chaves.

3.3.1 O índice invertido

Para essa indexação foi utilizada a ferramenta Lucene[9] e foram aplicados os seguintes tratamentos léxicos:

- Normalização: remoção de acentos, pontos e outros caracteres não importantes para as buscas.
- Tokenização: foi decidido realizar a tokenização por termos individuais, ou 1-gram. Foi verificado em testes que, para a busca no índice, o fato de se usar 2 ou mais n-grams não afetariam o resultado de busca por palavras chaves. Uma vez que ao buscar um laudo com os termos "forma habitual", o documento que apresentar estes termos próximos serão exibidos mais ao topo da lista.
- Remoção de termos insignificantes ("stopwords"): Foram utilizados os termos: a, ao, cuja, da, de, esse, este, entre outros listados na classe org.apache.lucene.analysis.br.BrazilianAnalyzer.getDefaultStopSet().
- Derivação e Lematização (Stemming and Lemmatization): foi utilizado somente a derivação, que consiste em eliminar plurais, prefixos e sufixos, deixando somente uma parte invariável do termo. Assim, os termos: **tireoide**, **tireoideano**, **tireoidite**, **tireoide**, **tireoidina** e **tireoideo** são indexados como '**tireoid**'.

Uma classe BasicAnalyzer, que estende a classe AbstractAnalyzer do Lucene, foi implementada para definir o pipeline de tratamento e aplicar as transformações descritas acima.

[9] http://lucene.apache.org/

3.3.2 Sinônimos e dicionário

Os laudos são redigidos por diferentes profissionais e uma mesma situação observada pode ser descrita usando termos diferentes, levando em conta a formação e experiência do profissional. Para minimizar o impacto desta situação nos resultados das buscas, um dicionário de sinônimos foi criado. Este dicionário é aplicado ao conteúdo do laudo antes da indexação e também na fase de busca.

Observou-se também, que muitos laudos apresentam erros de digitação, como por exemplo, foram encontradas ocorrências de 'nornal', 'cc', 'ñao', 'nao', 'naõ', 'cm.Volume' entre outros. Estes erros interferem no processo de indexação e busca, pois são termos não encontrados na língua portuguesa. Para resolver este problema foi utilizado o dicionário léxico Hunspell[10] com os arquivos para Português (pt_BR.aff e pt_BR.dic). Dessa forma, cada termo presente nos laudos é analisado pelo dicionário e caso não seja reconhecido, algumas sugestões para substituição são sugeridas. As substituições podem ser pré-definidas pelo operador no dicionário de sinônimos.

Um módulo foi implementado para auxiliar o operador a criar seus dicionários de sinônimos e substituições. Nesse módulo é possível visualizar o conteúdo dos laudos de um banco de dados e verificar sua corretude ortográfica. Os problemas identificados então podem ser tratados diretamente pelo operador na interface gráfica e as correções e substituições são armazenadas no dicionário para seu posterior uso durante as fases de indexação e busca. É possível criar um dicionário para cada especialidade de exame.

3.3.3 Resultado da busca por palavras-chaves

A indexação convencional possui um bom desempenho para busca por laudos que possuam determinados termos. Com a integração do dicionário de sinônimos, o resultado é melhorado, pois ao buscar os termos regular ou normal, o dicionário pode tratar os sinônimos, abrangendo no resultado tanto laudos que apresentarem o termo 'normal' quanto o termo 'regular', quando estes forem definidos como sinônimos.

O tempo de resposta das buscas também é bastante satisfatórios. Uma busca pelos termos 'mioma submucoso' em 3.000 laudos, demora cerca de 19 milissegundos e foram seleciona 9 laudos. Uma busca pelos termos 'mioma subceroso' sofre uma correção pelo dicionário e os termos submetidos à busca no índice ficam 'mioma subseroso'. Essa busca demorou cerca de 22 milissegundos e retornou 12 laudos.

As deficiências do índice convencional começam-se a ser notadas quando são submetidas buscas um pouco mais complexas, como por exemplo, quando são buscados termos que necessariamente devem estar relacionados a um determinado órgão ou expressam valores numéricos.

A busca pelos termos 'ovário direito contornos regulares' retornou 1.718 laudos em 83 milissegundos, dentre os 3000 indexados. O segundo laudo mais relevante da busca foi o seguinte:

[10] http://hunspell.sourceforge.net/

"Ultrassonografia Pélvica Transvaginal. Útero não identificado. Ovário Direito não identificado. Ovário Esquerdo tópico, contornos regulares e ecotextura com cisto de 1,4 cm. Dimensões 2, 5 x 2,1 x 2,0 cm. Volume estimado 5,9 cm. Impressão Diagnóstica. Útero ausente. Ovário esquerdo com formação cística."

Como pode ser visto, o laudo acima atende à busca por conter todos os termos. No entanto, o ovário direito nem se quer foi identificado durante esse exame. Logo, percebe-se o prejuízo da falta de uma análise semântica da busca, pois esperava-se como retorno somente laudos que apresentassem o ovário direito com contornos regulares.

O mesmo acontece com o resultado da busca 'ovário direito volume 5,9 cm', que inclui o seguinte laudo:

"Ultrassonografia Transvaginal. Bexiga vazia. Útero em Avf, com ecotextura miometrial sólida e homogênea, contornos regulares. Mede nos seus maiores eixos longitudinal, ântero-posterior e laterolateral respectivamente 5,9 x 2,5 x 4,1 cm. Volume uterino de 33,7 cm^3. Eco endometrial medindo 2,0 mm de espessura, sem alterações. Cavidade uterina virtual. Ovário direito Medindo 3,1 x 2,7 x 2,2 cm nos seus maiores eixos. Volume de 10,0 cm^3. Ecotextura heterogênea com cistos de 1,2 cm e 1,3 cm. Ovário esquerdo não identificado. Ausência de líquido livre na escavação retro uterina. Não evidenciam-se massas ou tumores nas regiões anexais. Conclusão. Ovário direito com formações císticas."

Buscas como 'ovário direito volume entre 2 e 6 cm' são impossíveis de serem satisfatoriamente atendidas em índices convencionais, uma vez que a definição de intervalos não será corretamente interpretada. Mesmo usando uma expressão específica para buscar valores numéricos no índice, a busca pelos valores não será atendida se o dado não estiver indexado apropriadamente para este fim. Mesmo assim, o valor numérico indexado perderia a ligação com o seu contexto, o que invalidaria novamente o resultado semântico da busca.

Tendo todas essas limitações bem compreendidas, o projeto seguiu com seu objetivo de analisar o conteúdo dos laudos e tentar estruturar as informações disponíveis.

3.4 Estudos e definições sobre ontologias

Com o uso de ontologia é possível expressar conhecimento, conceitos e sinônimos, além de permitir a realização de buscas por estes itens, mas a aplicação desta tecnologia não é trivial. Essa parte do projeto foi a mais difícil, pois são encontrados muitos materiais sobre a criação e busca por ontologia. Porém, a extração de informações de textos para a geração de regras ontológicas ainda não é uma área bem consolidada.

Inicialmente, uma coleção de 100 laudos de Ultrassonografia Transvaginal foi escolhida aleatoriamente. O conteúdo destes laudos foi analisado. Havia uma pergunta que precisava ser respondida: Como estruturar a informação que está distribuída neste conteúdo textual?

Foi observado no conteúdo dos laudos o relato de órgãos, patologias e várias características. Algumas características possuem valores nominais e outros valores numéricos. Era necessário armazenar todas essas informações sobre um determinado laudo em um formato que facilitasse futuras buscas. Então, para a extração do conhecimento sobre os órgãos e características observados nas imagens dos exames e anotados nos laudos foram definidas as seguintes classes ontológicas:

- ORGAN: Identifica um órgão observado. Exemplo: Ovário, Útero, etc.
- PATHOLOGY: Identifica uma patologia observada. Exemplo: Mioma, Cisto, etc.
- REGION: Indica a região onde o órgão foi encontrado: Direito, Esquerdo, Superior, Inferior, etc.
- FEATURE: Indica uma característica observada sobre um órgão ou sobre uma patologia. Exemplo: Volume, Ecotextura, Dimensão, etc.
- FEATURE_VALUE: Indica o valor de uma característica. Exemplo: Normal, Regular, Sólido, etc.
- FEATURE_MEASURE: Indica a unidade de medida na qual foi expresso um valor numérico de uma característica. Esta informação é utilizada para padronizar a escala dos valores, uma vez que algumas características podem estar expressas em cm, outras em mm, porém para a busca é importante encontrar o valor de uma característica independente de como ela foi registrada no laudo. Como por exemplo, ao buscar um ovário direito com volume entre 1 e 3 cm, a busca deverá retornar todos os ovários cujos volumes foram expressos entre 10 e 30 mm. Esta capacidade de lidar com diferentes unidades de medida, permitiu à ferramenta atingir um resultado nas buscas com maior confiabilidade e precisão.
- OTHERS: Esta classe é utilizada para os termos que não foram classificados pelo analisador. Ela é útil para verificar o desempenho do analisador e identificar possíveis termos úteis que não estão sendo corretamente classificados.

Como a linguagem utilizada nos laudos é muito especializada (descritiva e com poucos verbos), os analisadores gramaticais apresentam dificuldades para identificar sujeitos e predicados. Dessa forma, um analisador personalizado foi implementado para analisar o laudo sentença por sentença e identificar a presença de um órgão, uma patologia, uma região, uma característica, um valor de uma característica ou uma unidade de medida.

Para exemplificar o processo do analisador, dado o trecho do laudo:

"Ovário esquerdo, medindo 4,7 x 2,7 x 4,2 cm nos seus maiores eixos.

Volume de 26,6 cm^3. Cisto homogêneo de 35 x 23 mm."

Há três sentenças neste trecho. Cada sentença é analisada sequencialmente. Para cada sentença, uma ou mais regras ontológicas são instanciadas:

a) "Ovário esquerdo, medindo 4,7 x 2,7 x 4,2 cm nos seus maiores eixos."

ORGAN: ovário
REGION: esquerdo
PATHOLOGY: (vazio)
FEATURE: eixo
FEATURE_VALUES: 4,7 2,7 4,2
FEATURE_MEASURE: centímetro
OTHERS: medindo, x, x, em, o, seu, maior

b) "Volume de 26,6 cm³."

Nesta sentença não há referência de órgão ou patologia, somente uma característica volume é identificada. Nesta situação, o analisador mantém o contexto anterior e instancia uma regra para armazenar o volume de 26,6 cm³ do ovário esquerdo identificado na sentença anterior.

ORGAN: ovário
REGION: esquerdo
PATHOLOGY: (vazio)
FEATURE: volume
FEATURE_VALUES: 26,6
FEATURE_MEASURE: centímetro
OTHERS: de

c) "Cisto homogêneo de 35 x 23 mm."

Nesta sentença não há referência de órgão. O contexto anterior sobre o órgão e região são mantidos.

ORGAN: ovário
REGION: esquerdo
PATHOLOGY: cisto
FEATURE:
FEATURE_VALUES: homogêneo, 35, 23
FEATURE_MEASURE: milímetro
OTHERS: de, x

Cada regra recebe o identificador do laudo que a gerou. Quando uma sentença expressar mais de uma característica, o analisador extrai uma regra para cada característica. Como por exemplo, no trecho a seguir são referenciadas duas características (contorno e ecotextura) e dois valores (regular e homogênea):

"Útero (...) Os seus contornos são regulares e a ecotextura é homogênea"

As seguintes regras são instanciadas:

ORGAN: útero
REGION: (vazio)
PATHOLOGY: (vazio)

FEATURE: contorno
FEATURE_VALUES: regular
FEATURE_MEASURE: (vazio)
OTHERS: os, seus, são, e, a

ORGAN: útero
REGION: (vazio)
PATHOLOGY: (vazio)
FEATURE: ecotextura
FEATURE_VALUES: homogêneo
FEATURE_MEASURE: (vazio)
OTHERS: é

3.5 Indexação com ontologia

O conteúdo de cada laudo é analisado e várias regras ontológicas são extraídas. Para armazenar uma regra no índice, algumas otimizações são realizadas, principalmente para o tratamento de valores numéricos e unidades de medida. Todos os valores alfabéticos são armazenados em um campo FEATURE_VALUES no documento de indexação. Para os valores numéricos são utilizados os campos FEATURE_VALUE_NUM1, FEATURE_VALUE_NUM2 e FEATURE_VALUE_NUM3. Isto permite armazenar no máximo os valores de uma característica tridimensional, como por exemplo, a dimensão de um órgão que possui três medidas (eixo).

Além das informações textuais do laudo, alguns dados estruturados disponíveis na base de dados foram extraídos. Esses dados foram a idade do paciente, identificação da unidade de diagnóstico onde foi realizado o exame e a identificação do profissional que realizou o exame. Estes dados serão úteis para as futuras análises em conjunto com as características extraídas do conteúdo textual.

3.6 Busca

Um módulo específico para realização de buscas no índice foi implementado. A função deste módulo é permitir que o operador escreva uma frase de busca sem utilizar alguma linguagem técnica específica, utilizando somente expressões naturais encontradas nos laudos. A versão atual é bastante limitada, mas suficiente para obter os resultados preliminares do projeto. A Figura 8 mostra a tela onde a busca é realizada e os laudos selecionados são exibidos. Há uma opção para exportar o resultado da busca para uma planilha, permitindo que análises estatísticas sejam realizadas.

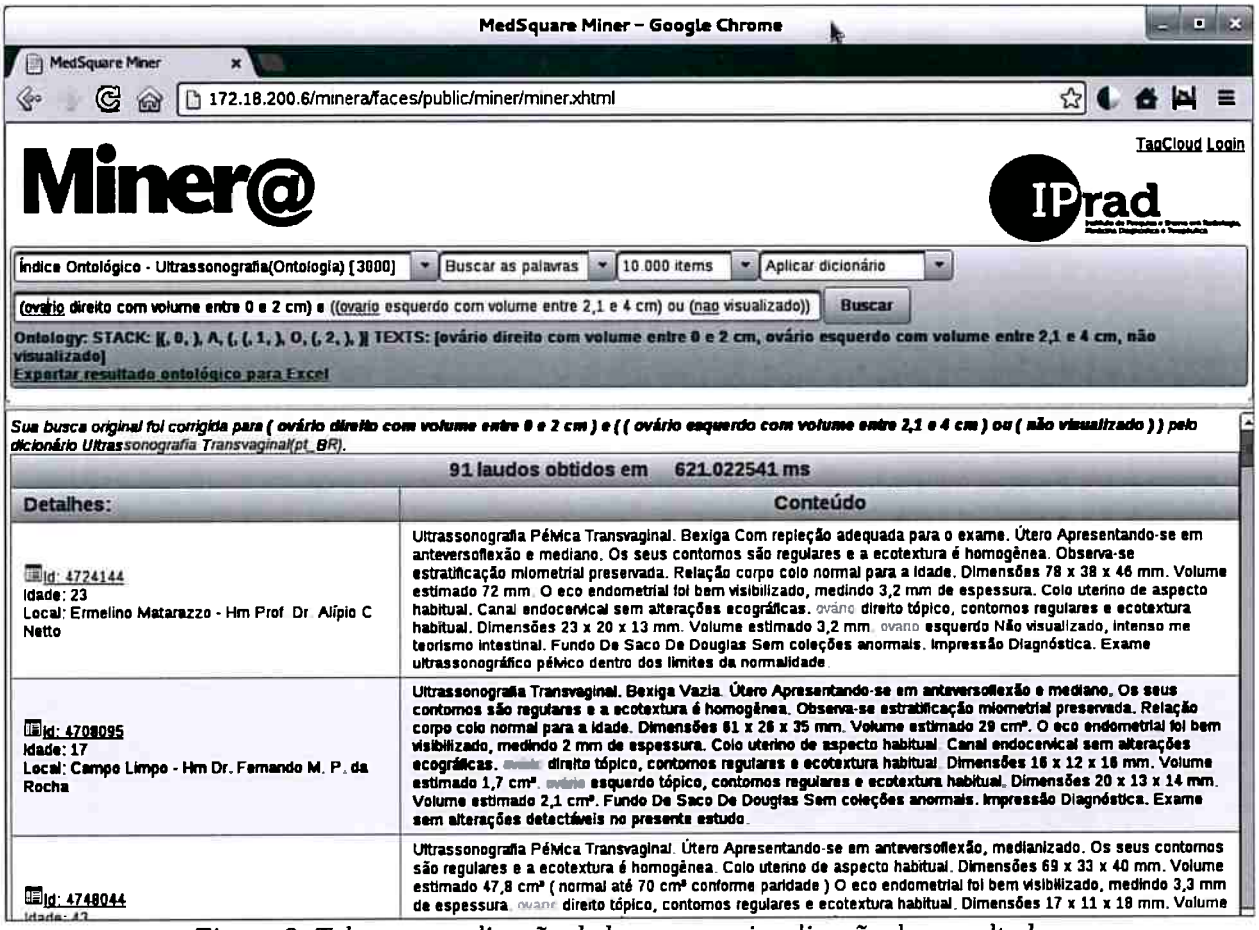

Figura 8: Tela para realização de buscas e a visualização dos resultados.

As frases de buscas são processadas por um interpretador semelhante ao analisador utilizado durante a indexação do laudo. É permitido o uso de parênteses e dos operadores lógicos E e OU. Assim, em uma mesma frase de busca podem ser fornecidas várias sentenças que devem ser separadas por ',' ou'.', ou estar entre parênteses.

Cada sentença é analisada individualmente para verificar as classes ontológicas que ela referencia. Então são criadas regras ontológicas que expressam o que o operador deseja buscar. O índice é consultado para verificar quais laudos satisfazem as regras encontradas. Para cada regra é obtido um conjunto de laudos. Esses conjuntos são operados conforme a organização parentética e os operadores lógicos.

A seguir são mostradas algumas frases de buscas e como o interpretador as interpreta:

"Encontrar ovário direito com volume de 22 cm³".

Esta frase cita um órgão **ovário** na região **direita** e a característica **volume** com valor numérico igual a **22** centímetros cúbicos. Os demais termos são desprezados. A seguinte regra de busca é então criada:

ORGAN: ovário
REGION: direito
PATHOLOGY: (Vazio)
FEATURE: volume
FEATURE_VALUES: 22
FEATURE_MEASURE: centímetro
OTHERS: encontrar, com, e, de

O índice então é consultado para verificar se alguma regra indexada é compatível com esta regra. Cada regra obtida tem a identificação do laudo que permite a recuperação do conteúdo original do laudo no banco de dados da ferramenta. Caso várias regras de um mesmo laudo sejam compatíveis com a regra de busca, o laudo é retornado somente uma vez, e vinculado com todas as suas regras compatíveis.

"Ovário direito com volume entre 10 e 22 cm³".

Esta frase apresenta um recurso interessante do módulo de busca, pois permite que o usuário defina um intervalo para valores numéricos de uma característica. O interpretador detecta o uso do termo 'entre' e coleta os próximos dois valores numéricos para criar um intervalo fechado. A seguinte regra de busca é gerada:

ORGAN: ovário
REGION: direito
PATHOLOGY: (Vazio)
FEATURE: volume
FEATURE_VALUES: [10, 22]
FEATURE_MEASURE: centímetro
OTHERS: com

"Ovário direito com volume entre 0 e 10 cm³, e ovário esquerdo com volume entre 10,1 e 20 cm³".

Esta frase apresenta duas sentenças separadas por vírgula e conectadas por um operador lógico E. O interpretador analisa as duas sentenças separadamente e sequencialmente. Para cada sentença uma regra de busca é gerada, como segue:

R1:
ORGAN: ovário
REGION: direito
PATHOLOGY: (Vazio)
FEATURE: volume
FEATURE_VALUES: [0, 10]
FEATURE_MEASURE: centímetro
OTHERS:

R2:
ORGAN: ovário
REGION: esquerdo
PATHOLOGY: (Vazio)
FEATURE: volume
FEATURE_VALUES: [10,1, 20]
FEATURE_MEASURE: centímetro
OTHERS:

O interpretador então obtém o conjunto CR1 de regras que satisfaz R1 e o conjunto CR2 de regras que satisfaz R2. Como cada regra identifica um laudo, são gerados os conjuntos de laudos CLR1 a partir de CR1 e o conjunto de laudos CLR2 a partir de CR2. Então é calculada a intersecção entre CLR1 e CLR2, uma vez que o operador lógico utilizado na frase de busca foi o operador E. Os laudos contidos no conjunto resultante da operação CLR1 ∩ CLR2 são obtidos no banco de dados.

A versão atual do interpretador também suporta frases de busca como:

- mioma intramural, ou mioma subseroso, ou mioma submucoso
- (ovário não identificado) ou (ovário não visualizado)
- (ovário direito com volume entre 0 e 2 cm^3) e (ovário esquerdo com volume entre 2,1 e 4 cm^3, ou não visualizado)

3.7 Resultados

Com o uso da ferramenta, a fundação FIDI conseguiu reunir em um único lugar os dados de laudos que se encontravam em diversas bases de dados. Durante o processo de indexação é possível construir diversos índices e selecionar quais categorias de exames deverão ser importadas para o índice. A Figura 9 mostra o módulo de importação de laudos.

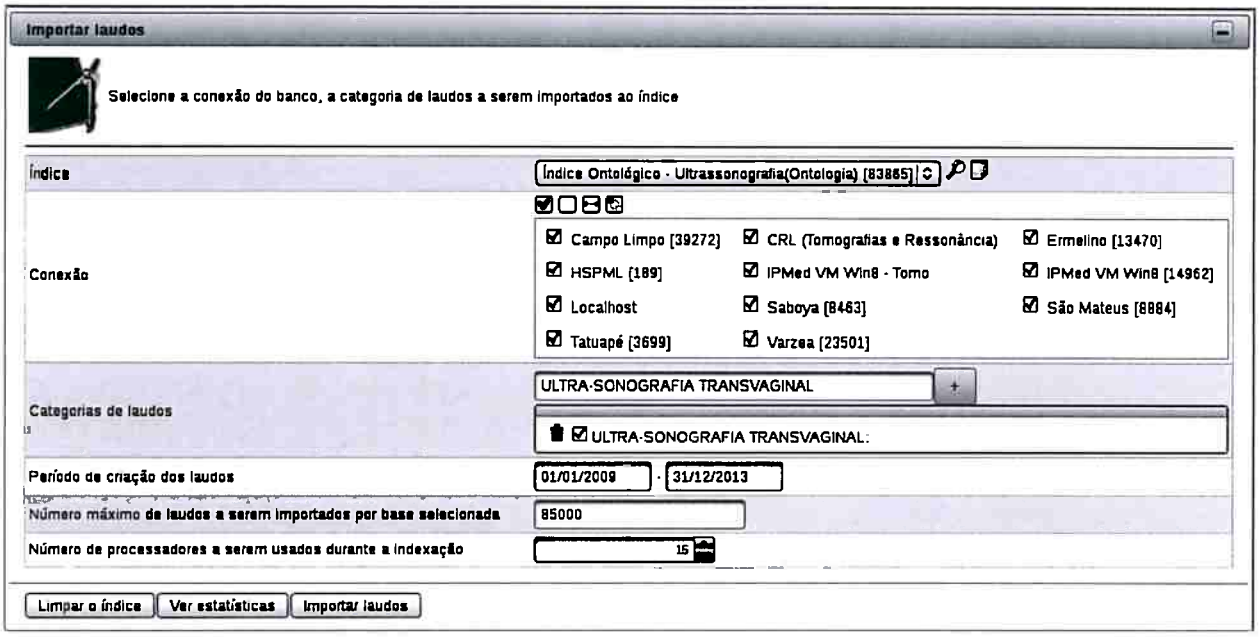

Figura 9: Módulo de importação de laudos.

Utilizando um servidor disponibilizado pelo setor de TI da empresa, foi configurada uma máquina virtual com sistema operacional Debian Wheezy 7.0 (64bits), com 16 núcleos, com 6 GB de memória RAM e dois volumes de armazenamento, um de 10 GB e outro de 100 GB.

Nesta máquina foram instalados os seguintes programas:

- Banco de dados MySql 5.5.31-0+wheezy1:
- Servidor de aplicações GlassFish Server Open Source Edition 3.1.1 (build 12)
- Máquina virtual Java OpenJDK 1.7.0_25
- Sistema de Versionamento CVS 1.12.13-MirDebian-9.

Com esta configuração, a ferramenta conseguiu indexar 83.865 laudos em aproximadamente 4 hora e 40 minutos, o que representa uma taxa de importação de aproximadamente 5 laudos por segundo.

Atualmente a ferramenta está sendo utilizada para selecionar casos de estudos e levantamentos estatísticos. A Figura 10 mostra um levantamento efetuado a partir de 7.001 laudos que apresentaram um dos ovários com volume entre 10,1 e 50 centímetros. A Figura 11 mostra um levantamento a partir de 586 laudos que apresentaram cisto ovariano com volume entre 3,1 e 100 centímetros. É importante ressaltar que, segundo a literatura médica, somente é considerado um cisto, quando o nódulo observado apresenta um volume maior que 3 centímetros. Ocorre que muitos especialistas acabam relatando nos laudos o termo cisto para nódulos com volume até 3 centímetros. Logo, esses casos são falsos positivos para cisto ovariano e não podem ser considerados em uma seleção de estudo.

Figura 10: Análise do volume do ovário por faixa etária.

Figura 11: Análise de casos de cisto ovariano por faixa etária.

4 Análise textual e estatística dos laudos

Os laudos são compostos em uma linguagem informal e bastante descritiva (). Esta característica facilita a extração de informações que

Para melhor caracterizar o escopo da aplicação da representação proposta, este capítulo apresenta um levantamento estatístico sobre as características dos textos dos laudos. Foi utilizada uma base com 4259 laudos de ultrassonografia transvaginal. Foram analisadas as frequências dos termos nos documentos e a frequência dos documentos em que o termo aparece.

Tabela 2: Lista dos termos com maior número de ocorrência absoluta

WORD	POS	Laudos	Ocorrência
NÚMERO	NÚMERO	3467	31600
cm	n	3825	10554
Volume	n	3918	7991
maiores	adj	3805	7878
seus	pron-det	3809	7858
eixos	n	3805	7848
uterina	adj	3992	7764
cm^3	n	3290	6575
normal	adj	3013	5491
Ovário	n	2594	5334
Útero	n	4072	4910
medindo	v-ger	3692	4115
Bexiga	v-fin	4096	4097
vazia	adj	4085	4085
Conclusão	prop	4084	4084
Ultrassonografia Transvaginal	n	4070	4070
uterino	adj	3872	4064
Medindo	prop	2264	4048
livre	adj	4014	4030
líquido	n	4010	4028
anexiais	adj	4011	4013
regiões	n	4011	4012

Tabela 3: Lista dos termos presentes no maior número de laudos

WORD	POS	n	sum
Bexiga	v-fin	4096	4097
vazia	adj	4085	4085
Conclusão	prop	4084	4084
Útero	n	4072	4910
Ultrassonografia Transvaginal	n	4070	4070
livre	adj	4014	4030
anexiais	adj	4011	4013
regiões	n	4011	4012
massas	n	4011	4011
líquido	n	4010	4028
tumores	n	4006	4007
uterina	adj	3992	7764
evidenciam	v-fin	3988	3988
Ausência	n	3972	4001
retro	adv	3971	3976
escavação	n	3970	3970
Volume	n	3918	7991
uterino	adj	3872	4064
cm	n	3825	10554
maiores	adj	3805	7878
eixos	n	3805	7848
mm	n	3787	4643

Tabela 4: Comparação de ocorrências de categorias gramaticais no corpus NILC, propósito geral, e no corpus de laudos

	NILC		Laudos			
Categoria gramatical	**Palavras**	**%**	**Palavras**	**%**	**Delta: Laudos → Nilc**	**Delta: Nilc → Laudos**
Substantivos	7113650	24,85%	138282	34,20%	**137,59%**	**72,68%**
Verbos		15,02%		7,05%	**46,93%**	**213,08%**
Adjectivos		6,44%		23,24%	**360,95%**	**27,70%**
Pronomes pessoais	469797	1,64%	12760	3,16%	**192,25%**	**52,02%**
Preposições	5298623	18,51%	51469	12,73%	**68,75%**	**145,44%**
Conjunções	1264416	4,42%	12510	3,09%	**70,03%**	**142,80%**
Advérbios	1455573	5,09%	11068	2,74%	**53,82%**	**185,80%**
Determinantes	5520746	19,29%	22023	5,45%	**28,24%**	**354,16%**
s	409265	1,43%	2209	0,55%	**38,20%**	**261,75%**
Numerais	949774	3,32%	31600	7,81%	**235,50%**	**42,46%**
	28622968	**100,00%**	**404386**	**100,00%**		

5 Projeto Anlaudos

Como discutido no capítulo anterior, diversos trabalhos propõem soluções para a extração de informações estruturadas de fontes textuais e apresentam altos graus de precisão. Contudo, observa-se que as propostas analisadas exigem um alto grau de conhecimento em PLN e do domínio analisado para se definir as regras de extração e a sequência de execução das tarefas. Algumas técnicas exigem a análise antecipada de todo o corpus para realizar a extração das informações. Diante dessas necessidades, a atual proposta apresenta um método que ofereça um baixo custo de configuração, evitando o acoplamento de fontes externas específicas de um domínio de informações, e que realiza sua análise a partir de um fluxo incremental de documentos, ou a partir de um pequeno conjunto amostral.

A ideia desta proposta de pesquisa nasceu em 2012, quando os membros do grupo de pesquisa foram convidados para participar de um projeto para extração de informações de laudos de ultrassonografia. Durante esse projeto foi desenvolvido um método capaz de identificar a referência de órgãos e patologias, e identificar suas relações com características nominais e numéricas. No **Capítulo 4** é apresentada uma descrição detalhada dos resultados preliminares desse projeto. O principal componente do método é um analisador que percorre o conteúdo do laudo para identificar as sentenças e extrair as informações, relacionando-as com os seus contextos. Contudo, ele é muito específico para o conjunto de laudos para o qual ele foi desenvolvido. Analisando a aplicação do analisador em laudos de outras modalidades de exames como radiografias, ressonâncias magnéticas e tomografias computadorizadas, observou-se que sua adaptação seria bastante custosa e os resultados seriam insatisfatórios devido às limitações presentes no método.

Uma pergunta foi então levantada: Como permitir, de uma forma simples, que um especialista médico interaja com o conteúdo textual e defina como as informações devem ser extraídas pela ferramenta?

A resposta não é tão simples, pois o conteúdo dos laudos apresentam, além dos desafios comuns do PLN, alguns desafios extras como erros ortográficos, incompletude linguística, domínio restrito e outros que serão destacados logo a seguir. De qualquer forma, a resposta foi buscada na própria pergunta: permitir que o usuário especialista simplesmente destaque as relações entre um termo e outro, e o método se encarrega das demais inferências.

Utilizando as ligações estabelecidas pelo usuário entre os termos no documento, o método deverá ser capaz de extrair algumas semânticas a partir de padrões de ligações, independentemente da linguagem utilizada no conteúdo do documento. A ideia inicial é que nenhuma informação adicional sobre o domínio seja adicionada ao método, simplificando sua implementação e diminuindo sua especialização em uma determinada modalidade médica. Contudo, alguma semântica para tratamento de números,

dimensões e escalas deverão ser adicionadas ao método para que ele seja capaz de interpretar os dados numéricos descritos nos laudos.

5.1 Descrição do método

O método consiste em representar o conteúdo de um laudo médico na forma de um grafo, definindo cada token do conteúdo como um nó, a princípio, cada nó está mapeado ao nó subsequente. Este é o fluxo original das palavras dentro do laudo. O usuário então inicia um processo de interação e associação dos nós, estabelecendo uma ligação vetorial entre eles. Esta ligação indica para o método que um determinado nó origem possui uma relação semântica qualquer com o nó destino. Ao final do processo de interação com o usuário, um grafo conceitual é construído para representar as ligações hierárquicas entre os nós, que são interpretados como conceitos. O método analisa no laudo original os nós precedentes, intermediários e subsequentes a cada conceito e gera características que são adicionadas aos rótulos das arestas do grafo conceitual.

As características das arestas são geradas a partir dos atributos léxicos e sintáticos de cada nó, como também, a partir de métricas que podem ser extraídas das ligações, ou seja, módulo, direção e sentido. Na análise dos nós intermediários, por exemplo, é possível verificar o número de nós que se encontram entre os dois nós das extremidades da ligação, o que pode indicar uma medida de distância entre os termos. Outras medidas poderão ser extraídas dessas ligações, à medida que se analisa as naturezas e combinações dos nós circunvizinhos.

Esse modelo baseado em grafo e arestas rotuladas deverá ser capaz de mapear os mesmos conceitos em outros conteúdos de laudos semelhantes. Assim, dado um grafo conceitual e um conjunto de laudos, será gerado, para cada laudo, um grafo resultante mapeando todos os conceitos encontrados. Com isto, será possível verificar quais laudos apresentam determinados conceitos associados. Ou ainda, quais laudos relatam determinados órgãos com determinadas características.

A Figura 12 mostra as três fases principais do método: ligação dos conceitos, montagem do grafo conceitual e extração das informações.

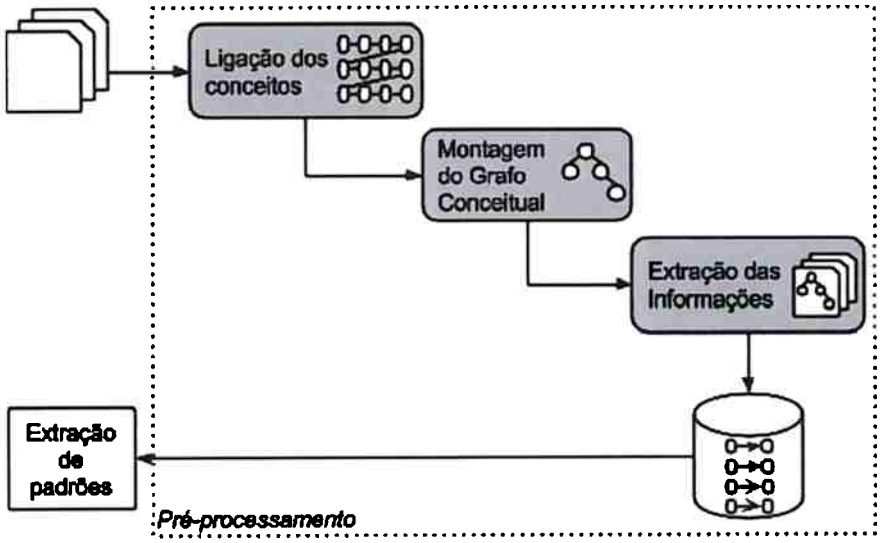

Figura 12: Fases do método proposto para a extração de informações estruturadas dos laudos médicos

5.2 Fase 1: Ligação dos conceitos

Nesta fase, um conjunto de laudos deverá ser selecionado para anotação dos conceitos. O critério de seleção dos laudos poderá ser manual ou automatizado pelo método. Para aumentar a relevância dos laudos sugeridos para uma anotação inicial é importante aplicar conceitos de variabilidade, como descrito no trabalho de Narciso (et al. 2011). Assim, de um conjunto de laudos muito semelhantes, somente um exemplar poderá ser selecionado para anotação, diminuindo o viés do modelo e aumentando sua eficácia na identificação de conceitos em laudos muito diferentes.

Após a seleção dos laudos, o usuário especialista deverá estabelecer ligações entre os termos, demonstrando que eles possuem alguma relação semântica. A Figura 13 mostra um documento com 10 termos. As arestas pontilhadas representam o fluxo original do documento. As arestas contínuas representam a ligação semântica entre os termos, estabelecida pelo usuário.

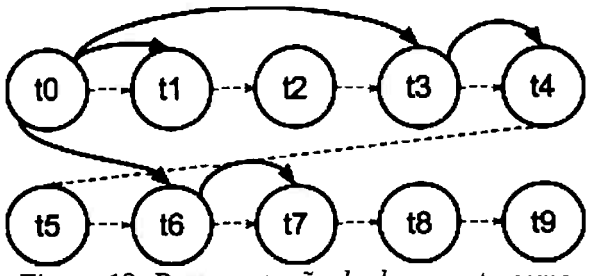

Figura 13: Representação do documento como um grafo e ligação dos conceitos

As ligações estabelecidas pelo usuário são elementos chaves para o método. Enquanto o método de Ciravegna (2001) usa um conjunto de documentos anotados para gerar e validar as regras de extração

de conceitos, a atual proposta usurá as ligações para extrair os conceitos e suas relações, sem que o usuário tenha que estabelecer delimitadores e suas relações.

5.3 Fase 2: Montagem do grafo conceitual

Esta é a fase mais crítica da proposta, pois nela é criado o modelo que será usado para extrair as informações nos demais laudos. A partir das ligações estabelecidas, um grafo conceitual deverá ser construído. Ele conterá todos os conceitos e relações destacados e servirá como modelo para a fase de extração de informações. A Figura 14 mostra um grafo conceitual baseado nas ligações da Figura 13. Os termos que não receberam ligações são descartados e não farão parte do modelo conceitual. Contudo, são importantes para a composição das estatísticas que serão agregadas ao modelo e anotadas como metadados nas arestas (relações).

Nesta proposta, a representação gráfica de um grafo conceitual difere da apresentada por Sowa (1984). No entanto, seus aspectos semânticos são ainda semelhantes, sendo os conceitos representados por círculos e suas relações por setas.

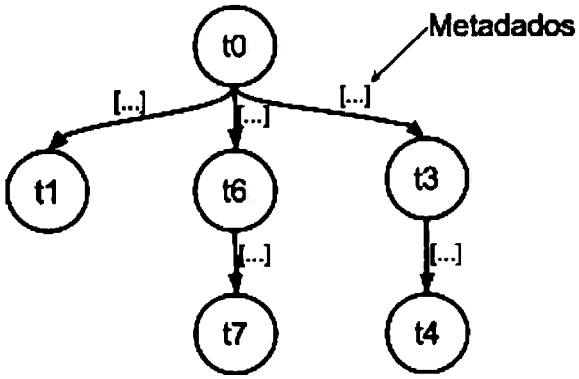

Figura 14: Grafo conceitual composto pelos termos do documento original.

O grafo conceitual poderá receber conceitos pré-definidos no modelo, como número, unidade de medida, negação e outros que serão discutidos na seção dos desafios do projeto. Além disso, a relação conceitual poderá ser realizada à medida que o usuário estabelece ligações entre dois termos, durante a fase 1. Isto permitirá ao modelo analisar os laudos que apresentam os mesmos termos em localizações semelhantes e gerar listas de laudos para validação do conceito em uma fase posterior ou em tempo real, enquanto o usuário ainda está anotando o conceito. Isto permitirá ao modelo aprimorar os metadados de cada relação, usando os casos positivos verdadeiros e falsos.

5.4 *Fase 3: Extração das informações em documentos semelhantes*

Nesta fase, o grafo conceitual servirá como base para a análise e extração dos conceitos presentes nos demais laudos. Os desafios encontrados nesta fase servirão de suporte para a identificação de requisitos para a fase anterior.

O processo de extração pode ser formalizado da seguinte maneira: dado um grafo conceitual G e um documento D, um grafo resultante R deverá ser construído tal que os nós de R representem instâncias dos conceitos de G encontrados em D, e as arestas de R representem as ligações dessas instâncias dentro de D, conforme as regras de G. A Figura 15 mostra o fluxo linear dos termos no documento D. A Figura 16 mostra um grafo conceitual G qualquer que determina o relacionamento entre seis conceitos. A Tabela 5 mostra o mapeamento dos conceitos de G e os termos do documento D.

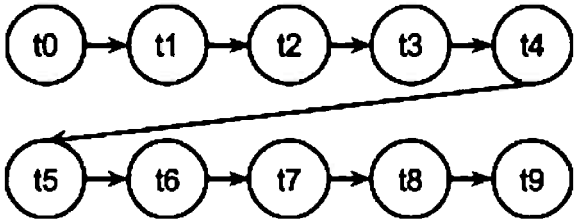

Figura 15: Representação do fluxo original dos termos no documento D.

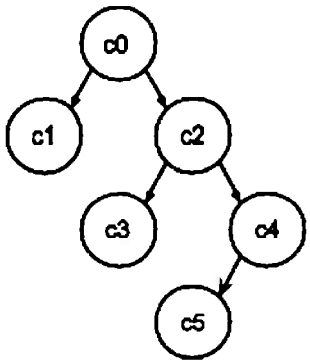

Figura 16: Grafo conceitual G.

Tabela 5: Mapeamento dos conceitos do grafo G e os termos do documento D.

Termo	Conceito
t0	c0
t1	c1
t2	c2
t3	c3
t4	*Desconhecido*
t5	*Desconhecido*
t6	*Desconhecido*

Termo	Conceito
t7	*Desconhecido*
t8	c4
t9	c5

O desafio é definir as arestas em R, que representam instâncias dos relacionamentos entre os conceitos de G dentro do documento D. No entanto, o documento D está escrito em linguagem natural, e pode conter diversos conceitos ainda não mapeados por G. Assim, poderão haver termos do documento D que pertencem aos conceitos em G, mas que localmente possuem uma relação semântica com outros termos ainda não mapeados. A Figura 17 mostra o resultado de R, desconsiderando a semântica dos termos t4, t5, t6 e t7 e assumindo que os termos t8 e t9 ainda sejam uma referência ao conceito c2.

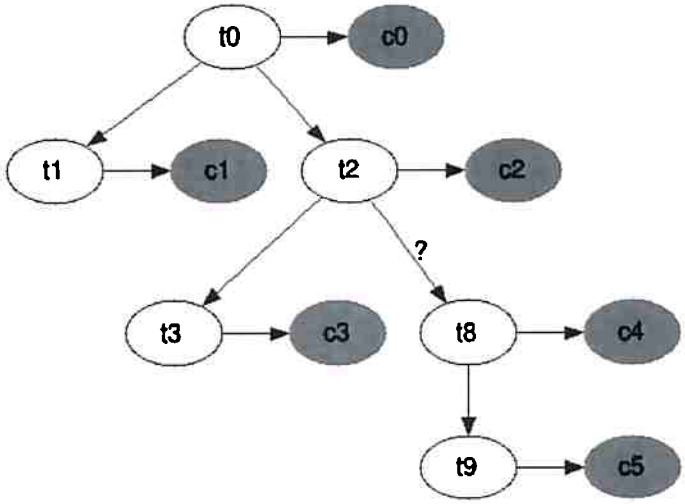

Figura 17: Grafo R com as instâncias dos conceitos de G encontradas no documento D.

Se um dos termos t4, t5, t6 ou t7 for um conceito ainda não mapeado em G, então os termos t8 e t9 podem ser referências a um conceito desconhecido por G e não ao conceito c2. Logo, esta instância de R será inválida, pois registra que no documento D foi encontrada uma ocorrência do conceito c2 ligado ao conceito c4.

Para melhor exemplificar o problema, as ilustrações a seguir aplicam o modelo descrito acima a um trecho simplificado de um laudo. A Figura 18 mostra o grafo original ligando os termos com seus subsequentes. A Figura 19 mostra um grafo conceitual com as relações entre 'ultrassonografia', 'transvaginal', 'ovário', 'direito', 'volume' e 'número'. A Tabela 6 mostra o mapeamento entre os conceitos do grafo G1 e os termos do documento D1.

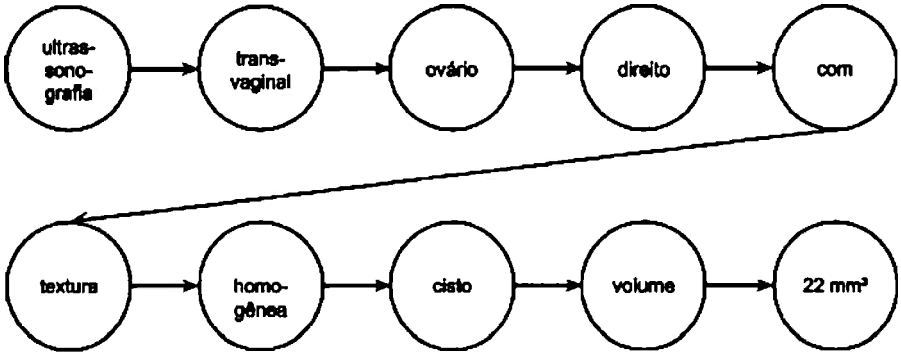

Figura 18: Representação do fluxo original dos termos no laudo D1.

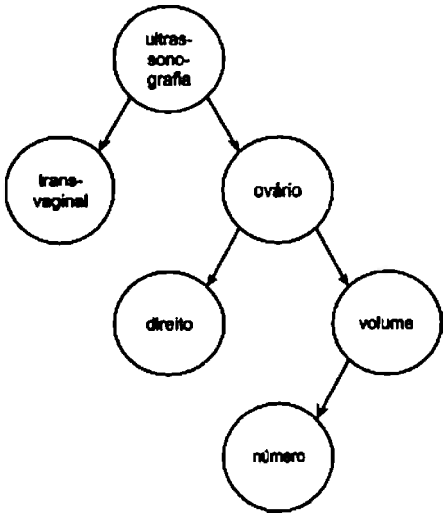

Figura 19: Grafo conceitual G1.

Tabela 6: Mapeamento dos conceitos do grafo G1 e os termos do laudo D1.

Termo	Conceito
ultrassonografia	ultrassonografia
transvaginal	transvaginal
ovário	ovário
direito	direito
com	*Desconhecido*
textura	*Desconhecido*
homogênea	*Desconhecido*
cisto	*Desconhecido*
volume	volume
22 mm³	número

É possível observar que, apesar de D1 conter todos os conceitos de G1, o conceito 'volume' não é referenciado para o conceito 'ovário', mas sim para o conceito 'cisto'. No entanto, como 'cisto' é um conceito desconhecido em G1, uma relação inválida no laudo D1 pode ser estabelecida entre 'ovário' e

'volume' se forem desconsiderados os termos existentes entre os dois conceitos. Uma forma de se evitar a associação inválida de conceitos é definir algumas características que delimitam essas associações durante a geração do grafo conceitual na fase anterior. Como por exemplo, adicionando ao grafo G1 uma característica de distância, ou seja, quantos termos existem entre os conceitos. Com isto, é possível atingir um maior grau de confiança ao estabelecer certas associações. A Figura 20 mostra em suas arestas uma característica 'd' que indica a distância encontrada entre os conceitos no momento da geração do grafo conceitual pelo usuário. Esta característica pode ser usada para validar a distância entre dois conceitos.

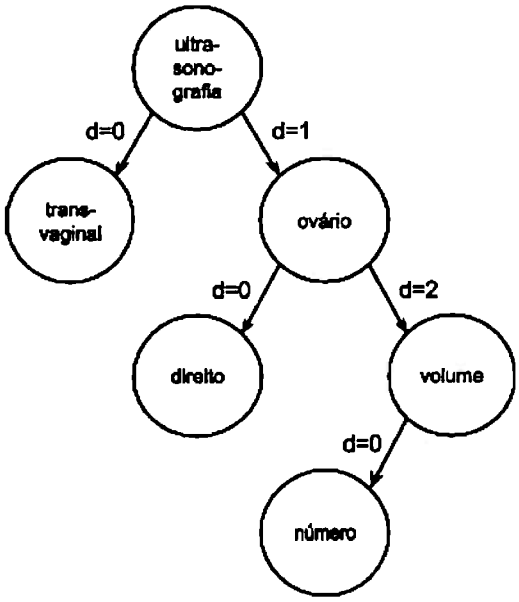

Figura 20: Grafo conceitual G1 com característica 'd' (distância) nas arestas que ligam os conceitos.

Aplicando a restrição da característica 'd' para relacionar os conceitos em D1, é possível definir com um alto grau de confiança que o conceito 'transvaginal' está associado ao conceito 'ultrassonografia', pois tanto no grafo conceitual G1 (Figura 20) quanto no laudo D1 (Figura 18), ambos conceitos não apresentam termos intermediários, ou seja, d=0. O mesmo ocorre com os conceitos 'ovário' e 'direito' que tendem a aparecer juntos nos laudos.

Em uma análise preliminar, é possível afirmar que todos os conceitos associados com d=0, tanto no grafo conceitual como no documento analisado, gerarão instâncias dessa associação no grafo R. Porém, o mesmo não pode ser afirmado quando d é diferente de zero em uma das fontes. Esta afirmação deverá ser verificada com experimentos que demonstrem relações conceituais com d=0 em todo o corpus.

Voltando à uma análise superficial da contribuição da característica 'd' na remoção de associações inválidas, o grafo G1 (Figura 17) mostra o mapeamento do conceito 'ovário' para 'volume' com uma característica d=2. Já no laudo D1 (Figura 18) os conceitos 'ovário' e 'volume' estão separados por cinco termos, ou seja, d=5. Logo, o grau de confiança pode ser descrito com uma função de

probabilidade que pode ser estimada através da frequência das ligações usando a forma interativa de treinamento:

```
grauDeConfiança(termo1, termo2, D, G):
    sendo c1 = conceito(termo1, G)
        e c2 = conceito(termo2, G)
        e d1 = distância(termo1, termo2, D)
        e d2 = metadadoDistância(c1, c2, G);
    se d1 = d2 = 0;
    então grauDeConfiança = 1;
    senão grauDeConfiança = menor(|d1|, |d2|)/maior(|d1|, |d2|).
```

Esta é somente uma suposição inicial que será verificada e evoluída com o andamento das pesquisas. Até porque outras características serão exploradas e analisadas para verificar sua influência no grau de confiança da associação entre dois conceitos quaisquer existentes em G e identificados em D. Além disso, poderão ser utilizadas técnicas de aprendizagem de máquina para se definir uma função de regressão que determinará o grau de confiança baseando em várias características. Contudo, neste caso, deverá ser verificado como reduzir o viés e a variação da função tendo poucos exemplares anotados para treinamento e teste.

A seguir é descrito um pseudo algoritmo para a extração das relações entre os conceitos presentes em um laudo, dado um grafo conceitual contendo os metadados necessários para a função **grauDeConfiança()**.

```
grafoResultante(Laudo l, GrafoConceitual gc):
   sendo gol = grafoFluxoOriginal(l)
       e grl = grafoResultanteVazio();
   para cada termo t1 em gol faça:
      se t1 pertence aos conceitos de gc então:
         adiciona t1 em grl;
         adiciona conceito(t1) em grl;
         liga t1 ao conceito(t1) em grl;
         para cada termo t2 seguinte a t1 em gol faça:
            se t2 pertence aos conceitos de gc então:
               adiciona t2 em grl;
               adiciona conceito(t2) em grl;
               liga t2 ao conceito(t2) em grl;
               se grauDeConfiança(t1, t2, gol, gc) > LIMIAR então
                  liga t1 a t2 em grl;
   então grafoResultante = grl;
```

Outra função importante para o método será a função **conceito()** que determina a qual conceito um termo do laudo se refere. No entanto, os conceitos são marcados na **Fase 1**, pelo especialista, e como será discutido a seguir, há muitos desafios na normalização destes conceitos. As próximas seções destacarão os desafios enfrentados no tratamento do conteúdo dos laudos médicos e algumas abordagens

para superar cada um desses desafios. Um exemplo passo a passo será utilizado para destacar as situações reais que são encontradas nos laudos.

5.5 Os desafios para o método

Abaixo é mostrado um laudo de ultrassonografia. A seguir, na Figura 21, são mostradas algumas associações possíveis entre os termos do laudo. Os retângulos e as setas imprecisas foram utilizados para ilustrar a ideia que o método traz sobre uma anotação livre, realizada pelo especialista. As associações são direcionais e indicam que um determinado termo possui um relacionamento com outro termo. A semântica deste relacionamento é conhecida pelo especialista que realizou a associação, porém, para o método, esta semântica não será considerada. A formatação dos parágrafos do laudo foi removida para simplificar a sua representação.

ULTRASSONOGRAFIA TRANSVAGINAL Bexiga vazia. Útero visualizado (histerectomia sub-total). O colo mede: 3,1 x 3,0 x 1,8 cm. Ovário direito: Medindo 3,1 x 2,2 x 2,3 cm nos seus maiores eixos. Volume de 3,4 cm³. Aprsentando uma imagem cistica, de aspecto simples, medindo 21 mm (funcional?). Ovário esquerdo: nao visualizado (grande interposicao gasosa). Ausência de líquido livre na escavação retro uterina. Não evidenciam-se massas ou tumores nas regiões anexiais. CONCLUSÃO Cisto em ovario direito.

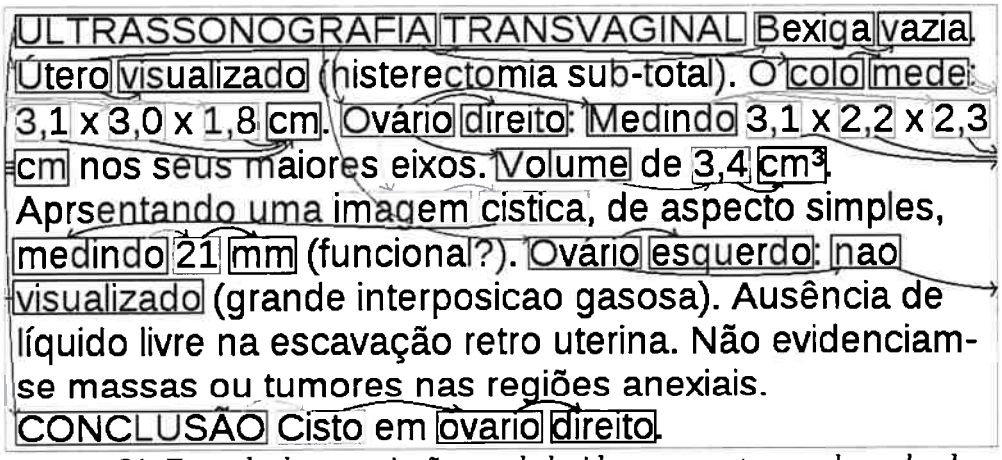

Figura 21: Exemplo das associações estabelecidas entre os termos de um laudo pelo médico especialista

Durante a ligação dos termos é importante que artigos, preposições, pronome, conjunção e interjeição não sejam destacados como conceitos, pois estes são componentes com papéis específicos para a linguagem, mas não são propriamente um conceito do domínio. No grafo conceitual, as arestas realizam o papel de associação entre os conceitos. A Figura 22 mostra o grafo obtido a partir das ligações estabelecidas entre os conceitos destacados no laudo.

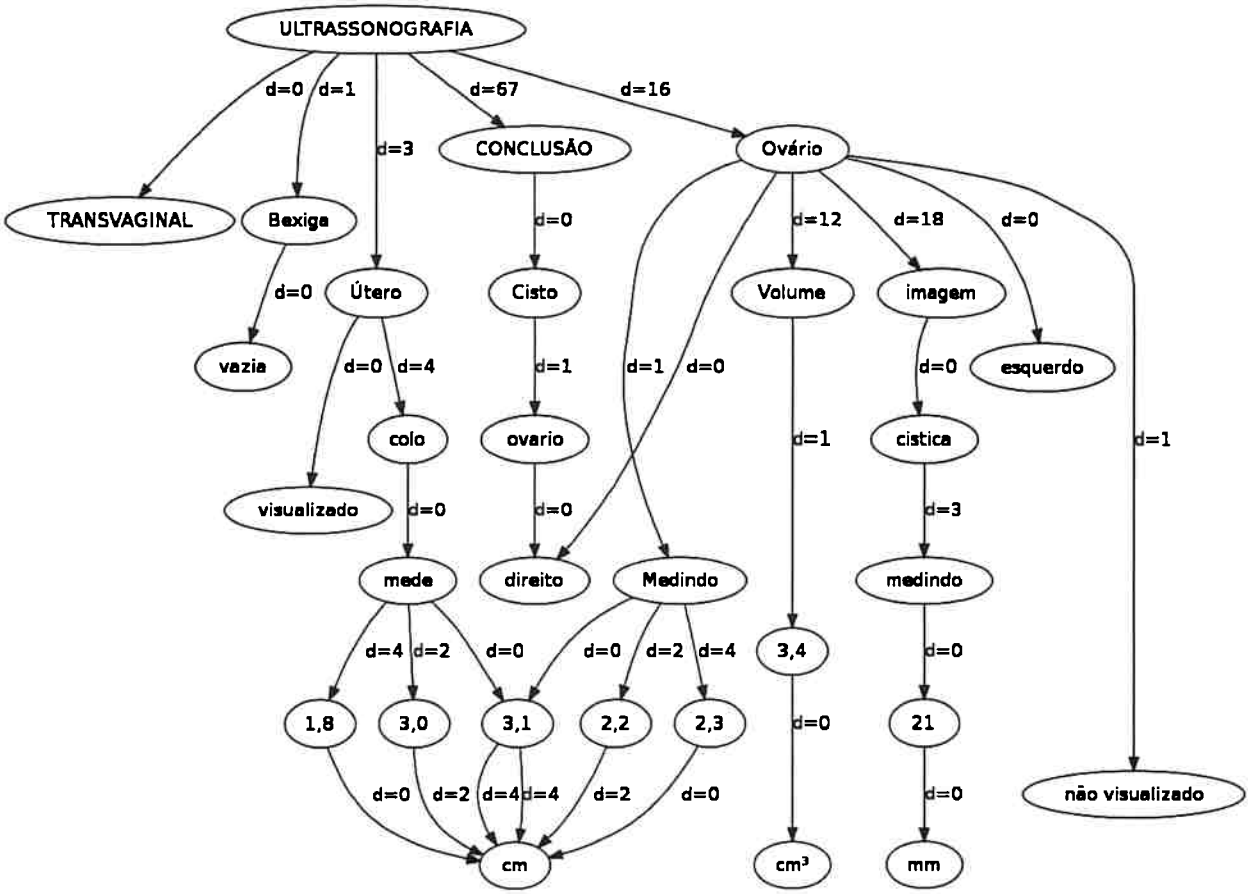

Figura 22: Arestas mostrando a distância (D) entre um par de nós. Para o cálculo de D foram considerados somente os tokens alfanuméricos.

É possível verificar visualmente na Figura 22 que o laudo é de uma ultrassonografia transvaginal e que relata os aspectos da bexiga, do útero e do ovário. Além disso, há o relato de uma 'CONCLUSÃO' que foi associada a um cisto no ovário direito. É possível extrair dados estatísticos sobre os conceitos destacados e os demais termos, sem considerar quaisquer etiquetas léxicas, sintáticas ou semânticas que possam ser associadas aos termos do documento. Nesse caso, foi extraída a característica 'd' que indica a quantidade de termos encontrados entre dois conceitos destacados no grafo. O desafio agora é identificar estes mesmos conceitos em conteúdos similares, mantendo a consistência semântica dos relacionamentos.

Com este exemplo inicial já é possível notar alguns desafios que deverão ser superados pelo modelo proposto, como a corretude ortográfica dos laudos, o tratamento de sinônimos, a falta da completude gramatical, o tratamento de valores numéricos, a identificação de valores multidimensionais, a identificação de valores expressos em intervalos, a representação de sentença negativa, o compartilhamento de valores e a relação do conceito com o contexto. As seções que se seguem discutirão cada desafio e o impactado de possíveis soluções nos objetivos do método proposto.

5.5.1 Corretude ortográfica

No laudo mostrado anteriormente verifica-se alguns problemas ortográficos que são listados na Tabela 7.

Tabela 7: Erros ortográficos encontrados em laudos de ultrassonografia transvaginal e sugestões de correção

Erro	Correção
sub-total	subtotal
Aprsentando	Apresentando
cistica	cística
nao	não
interposicao	interposição
ovario	ovário

Analisando outros laudos, foram encontradas, por exemplo, outras combinações para palavra não: ñao, naõ e ãno. Serapião (et al., 2010) verificou que em um conjunto de 22.247 laudos de mamografia, foram identificados 4.435 termos diferentes, sendo que 934 (21%) estavam com a grafia incorreta.

Estes erros ortográficos fazem com que cada variação incorreta de um termo seja interpretado como um novo termo no conteúdo, ou seja, um novo conceito, e consequentemente um novo nó no grafo. O grafo da Figura 22 mostra o nó 'Cisto' ligado ao nó 'ovario', sendo que ele deveria estar ligado ao nó 'Ovário', que possui o mesmo significado.

Um dicionário pode ser utilizado para a correção de erros ortográficos. Contudo, os laudos possuem alguns termos muito específicos que podem não ser encontrados em um dicionário de propósito geral. Honorato (2008) propõe um método que analisa todos os termos de um conjunto de documentos e verifica a distância de edição entre os demais termos, usando a distância de Levenshtein (Kruskal, 1983). Um termo com baixa frequência e que possui uma grafia muito semelhante a de outros termos mais frequentes pode indicar um possível erro ortográfico dentro do corpus. Neste caso, os termos semelhantes compõem uma lista de possíveis correções. A interação do especialista é necessária para definir a correção mais apropriada. Não fica claro no trabalho como a ocorrência de plurais é tratada, uma vez que a distância de edição entre um termo no singular e o mesmo termo no plural será, geralmente, de uma inserção, ou seja, uma distância bem pequena. Isto causará a sugestão de erros para todas as ocorrências no plural ou no singular, dependendo de qual termo tiver maior número de ocorrências.

Na atual proposta, um grafo conceitual com os termos corretos é gerado a partir da interação do especialista com o conteúdo dos laudos. Partindo desta premissa, é possível analisar termos semelhantes,

ou seja, com uma pequena distância de edição, para identificar outras possíveis grafias encontradas no corpus. Esta abordagem manteria a independência de dicionários de uma linguagem específica.

Uma outra forma de automatizar o processo de correção é usar um corretor ortográfico para identificar os erros e sugerir correções para os termos mais gerais da linguagem. Para a atual proposta, o corretor Hunspell[11] está sendo analisado como uma opção, pois é um dicionário de código aberto e que oferece suporte para o português e outras línguas. Este dicionário é capaz de gerar uma lista de sugestões para a substituição ao encontrar um termo não reconhecido. Contudo, uma análise deve ser realizada para verificar o índice de acerto do primeiro item da lista de sugestões em um corpus de laudos médicos.

A adição de termos do domínio médico ao dicionário diminuiria as indicações de erros, porém, implicaria na integração de alguma fonte externa específica ao domínio do conteúdo. Outra solução é verificar a distância de edição entre o termo incorreto e os termos sugeridos, principalmente a primeira sugestão. Se o termo sugerido for muito próximo do original, um candidato à substituição foi encontrado. Caso contrário, o termo pode estar correto e não ser reconhecido pelo dicionário. Testes deverão verificar a eficácia desta abordagem para automatizar a substituição dos erros ortográficos.

Experimentos deverão ser realizados para verificar a eficiência e a eficácia das abordagens descritas anteriormente. Além disso, o uso de um dicionário faria com que o método ficasse dependente das línguas suportadas pelo dicionário e aumentaria a complexidade de implantação do método. O acoplamento de um dicionário também poderá ser opcional, ou seja, o método deverá estar preparado para trabalhar com um dicionário ou sem ele.

5.5.2 Tratamento de sinônimos

O grafo da Figura 22 mostra três diferentes representações para o conceito medir: 'mede', 'medindo' e 'Medindo'. Este é um exemplo de sinônimos próximos que podem ser facilmente identificados aplicando duas técnicas relativamente simples. A primeira é ignorar o caso das letras dos termos. Assim, 'Medindo' e 'medindo' serão identificados como sendo o mesmo conceito. A segunda técnica é aplicar a lematização aos termos e representar o conceito em sua forma canônica. Assim, os conceitos, 'mede', 'medição', 'medindo' e outras variantes serão identificados como sendo o conceito 'medir'.

No entanto, existem sinônimos mais complexos que não são resolvidos com lematização, como por exemplo:

i) "Útero não identificado" == "Útero não caracterizado"

ii) "Ovário direito não visualizado" == "Ovário direito ausente"

iii) "Cisto em ovário direito" == "Ovário direito com formações císticas"

[11] http://hunspell.sourceforge.net/

Os dois primeiros exemplos parecem mostrar sinônimos em uma linguagem mais geral e o terceiro exemplo mostra termos específicos do domínio médico. Tomando o primeiro exemplo e analisando as traduções dos termos 'identificado' e 'caracterizado' na WordNet.Pr[12], identified e characterized, nenhuma relação semântica é encontrada. O que leva à conclusão que a utilização destes termos como sinônimos foi definida, de forma particular, pelos especialistas da área médica. Logo, esta relação só será identificada a partir de uma ontologia específica que a descrevesse. O mesmo acontece com os sinônimos 'ausente' e 'não visualizados', que não são suportados pela WordNet.Pr. O termo 'não existente' foi o sinônimo mais próximo para 'ausente'.

Zhou (2005) usa uma abordagem manual para declaração de sinônimos. Um ano depois, Zhou (2006) acopla uma ontologia médica para identificar os possíveis sinônimos de um determinado termo. A primeira abordagem exige um esforço dos especialistas na criação da lista dos sinônimos que serão encontrados nos documentos e a segunda abordagem exige o acoplamento de uma ontologia específica para o domínio do conteúdo dos laudos.

A atual proposta tentará manter o seu objetivo de baixo acoplamento com fontes externas de informações semânticas específicas para um determinado domínio ou especialidade médica. No entanto, duas situações deverão ser investigadas:

a) Como associar sinônimos de propósito geral da linguagem, como: doença e patologia?

Uma possível resposta é a utilização de um tesauro de propósito geral que contenha a relação de sinônimo entre seus termos. Esta abordagem dependerá da qualidade e abrangência do tesauro utilizado. Além disso, como discutido acima, pode haver uma redefinição dos sentidos dos sinônimos em determinados domínios.

b) Como associar sinônimos mais específicos, como: não identificado, não caracterizado, não visualizado, ausente?

Para esta pergunta, pelo menos duas abordagens deverão ser analisadas:

A primeira é a possibilidade da marcação voluntária de sinônimos no momento em que o usuário destaca um determinado termo.

A segunda abordagem é a utilização de técnicas estatísticas para inferir sinônimos a partir da análise do corpus. Por exemplo, é possível identificar os termos precedentes e subsequentes de um determinado termo e verificar, no restante do corpus, quais termos aparecem entre esses termos. A mesma técnica pode ser aplicada utilizando somente os termos precedentes ou subsequentes. De qualquer forma, o termo encontrado será somente um candidato a sinônimo. As sentenças abaixo exemplifica o resultado não esperado da técnica descrita:

i) "... ovário **direito** não..." ii) "... ovário **esquerdo** não..."

[12] http://wordnetweb.princeton.edu/perl/webwn

Destacando o termo **direito** na primeira sentença, é possível obter o termo 'ovário' como precedente e o termo 'não' como subsequente. Analisando a segunda sentença, verifica-se que o termo **esquerdo** se encontra entre os mesmos termos da primeira sentença. Neste caso, o termo **esquerdo** poderá ser sugerido como um possível sinônimo para o termo **direito**. No entanto, os termos representam características distintas. Assim, a técnica não é determinística, porém é útil para gerar uma lista de possíveis sinônimos. Além disso, nenhum recurso específico para a língua portuguesa é necessário, uma vez que são utilizados somente dados estatísticos do próprio corpus.

A Figura 23 mostra o grafo resultante após a alteração do caso das letras para minúsculas e o tratamento de alguns sinônimos. Foram efetuadas as substituições de 'ovario' pela sua forma correta 'ovário', de 'imagem cistica' por 'cisto', e de 'medindo' e 'mede' por 'medir'. Os nós em cinza foram mantidos para lembrar quais termos foram substituídos. O resultado é um grafo com menor redundância de conceitos.

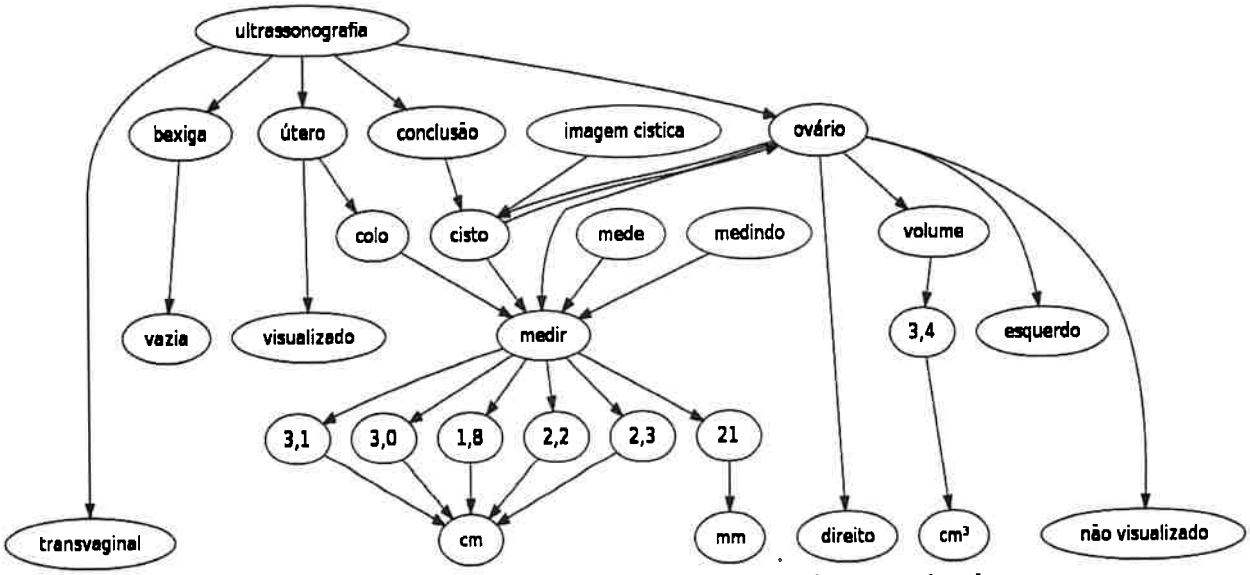

Figura 23: Grafo com os conceitos escritos em letra minúscula.

5.5.3 Completude gramatical

A linguagem utilizada em laudos é direta e não utiliza, em muitas vezes, as estruturas cultas. Dessa forma, observa-se a omissão de verbos e outros complementos linguísticos que trariam uma maior completude no sentido das sentenças. Algumas abordagens como Ding (2003) e Mashyastha (2003) usam técnicas para identificação de sujeitos, predicados e relações entre objetos diretos e indiretos. Essas técnicas conseguiriam interpretar corretamente uma sentença como:

"O volume do ovário direito é de 3,2 cm³."

Porém, como essas técnicas são dependentes das estruturas linguísticas, teriam dificuldades de identificar as relações em uma sentença como:

"Ovário direito: Medindo 3,1 x 2,2 x 2,3 cm nos seus maiores eixos. Volume de 3,4 cm³."

Zhou (2003) propõe uma abordagem híbrida que tenta identificar os elementos linguísticos presentes na sentença e estabelecer suas relações, porém, se ela falhar, alguns padrões são utilizados para tentar extrair as informações. Como por exemplo o padrão abaixo identificaria o conceito **Medindo** entre "direito:" e "3,1":

<substantivo>: * <número>

A atual proposta pretende extrair as informações utilizando o grafo conceitual construído pelo especialista, mantendo assim, sua simplicidade e independência de ferramentas mais complexas e dependentes de domínio. A completude gramatical do conteúdo não será importante. Melhor ainda, o método tentará se mostrar independente dela e facilmente aplicado a conteúdos com o mínimo de complementos linguísticos e pouco estruturados.

Contudo, as etiquetas léxicas e sintáticas podem fornecer alguns dados importantes sobre a estrutura do texto, os termos precedentes, intermediários e subsequentes. Estes recursos linguísticos serão analisados para verificar sua influência na precisão do método. O trabalho de Aires (2000) apresenta alguns etiquetadores clássicos, incluindo alguns já treinados para o Português do Brasil que serão avaliados.

5.5.4 Tratamento de valores numéricos

Além da discriminação de órgãos e patologias, os laudos médicos registram algumas características fisiológicas e patológicas através de valores numéricos. Muitos valores são associados às suas respectivas unidades de medida. Como por exemplo, a pressão arterial é expressa em termos da pressão sistólica sobre a pressão diastólica, e é medida em milímetros de mercúrio (mmHg). O pulso pode ser medido em batimentos por minuto (bpm). Já o peso pode ser registrado em quilograma, em libra, em onça, e etc. A temperatura corporal pode ser medida em graus Celsius (C°) ou graus Fahrenheit (F°). O volume pode estar em milímetros cúbicos (mm^3), em centímetros cúbicos (cm^3) ou em metros cúbicos (m^3), ou ainda, em polegadas cúbicas. Estes são somente alguns exemplos que demonstram que a extração de valores numéricos dependem diretamente da identificação da unidade de medida em que o valor está representado.

Alguns trabalhos como os de Zhou (2006) e de Honorato (2008) assumem que os valores numéricos das características extraídas do texto estarão sempre na mesma unidade de medida. Essa abordagem não seria um problema para valores que realmente tendem a ser expressos usando as unidades de medidas regionais. No entanto, analisando alguns laudos de ultrassonografia, observou-se que características como o volume e a dimensão de órgãos podem estar expressas tanto em mm^3 quanto em

cm^3, e que a área de um cisto, por exemplo, pode estar em mm^2 ou cm^2, dependendo muitas vezes do profissional que realizou a anotação ou do montante do valor observado.

Em um trabalho preliminar, descrito com mais detalhes no **Capítulo 4**, foi implementado um analisador de laudos que identifica os *tokens* referentes às unidades de medidas metro, centímetro e milímetro. O valor referenciado é padronizado para milímetros antes do seu armazenamento. O mesmo processo é aplicado durante a busca. Desta forma, se o usuário expressar valores em centímetros, a busca será efetivamente realizada em milímetros. Com isto, todos os valores originalmente definidos em metro, centímetro ou milímetro poderão ser recuperados por uma pesquisa realizada em qualquer uma destas unidades de medida. Esta abordagem mostrou-se eficiente, mas foi necessário acoplar ao mecanismo de indexação as definições e relações das unidades de medidas. A Figura 24 mostra um exemplo do processo de padronização da escala dos valores numéricos aplicado antes da indexação e da busca.

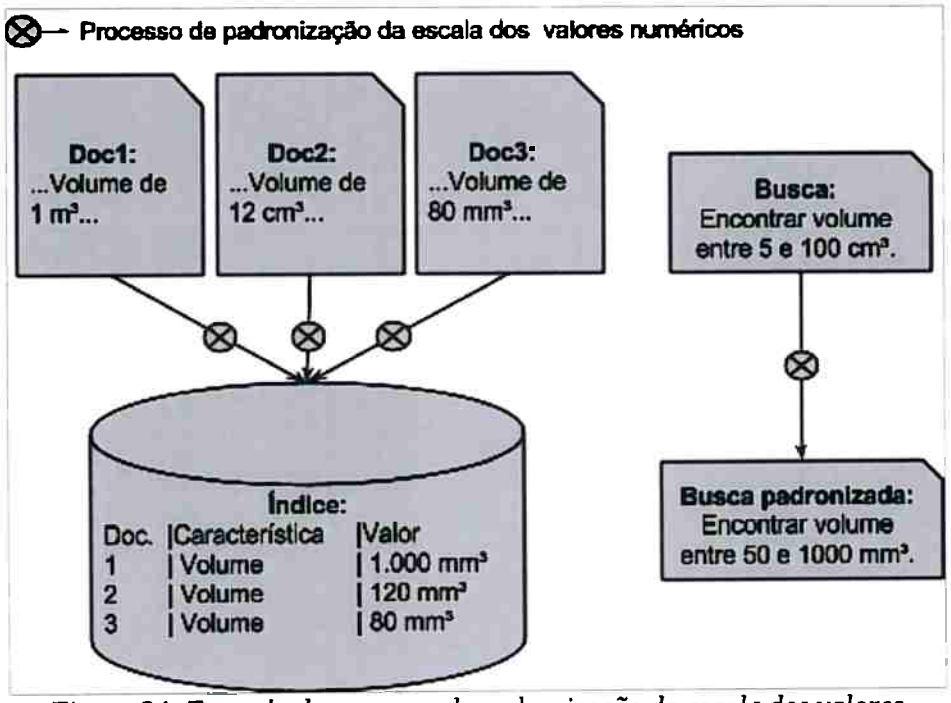

Figura 24: Exemplo do processo de padronização da escala dos valores numéricos para a indexação e busca.

Apesar de pouco comum nos laudos, os valores numéricos também podem estar registrados por extenso, e é importante que o modelo seja capaz de tratá-los. As ferramentas geradoras de etiquetas léxicas (*POS-taggers*) são capazes de identificar numerais expressos tanto por extenso quanto por dígitos, porém são específicas para uma língua.

Apesar do Brasil, Estados Unidos e outros 54 países fazerem partes do Bureau Internacional de Pesos e Medidas (BIPM, 2006), é necessário verificar se os laudos médicos obedecem ao Sistema Internacional de Unidades (SI). Ou seja, analisar a influência do SI na formação dos médicos especialistas que escrevem os laudos. Isto será importante para definir a parte do modelo que atuará sobre as unidades de medidas, principalmente no que se refere à sua internacionalização.

Acoplando o tratamento de valores numéricos ao método, os nós numéricos e de unidades de medidas pelo são substituídos pelos seus respectivos conceitos, como mostra a Figura 25.

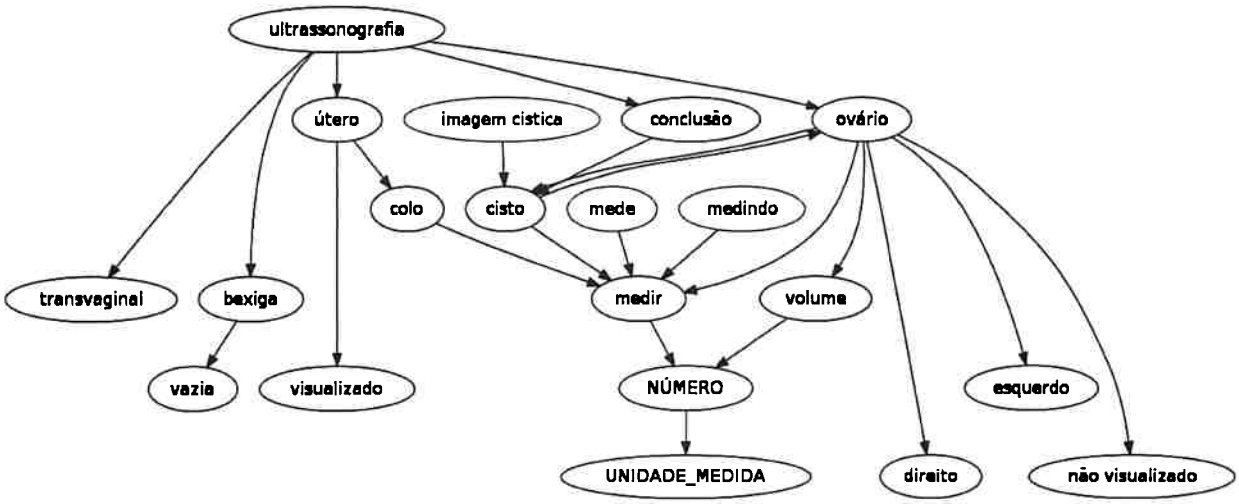

Figura 25: Grafo com os nós numéricos e de unidade de medidas substituídos pelos seus conceitos.

5.5.5 Identificação de valores multidimensionais

Algumas características numéricas encontradas nos laudos podem ser unidimensionais, bidimensionais ou tridimensionais, dependendo da técnica utilizada para a aquisição das imagens do exame. Aparentemente, a identificação das dimensões de um conceito pode ser uma tarefa com um certo grau de complexidade. No entanto, analisando a interação do usuário ao destacar as ligações no laudo, é possível verificar se um determinado nó possui uma ou mais arestas de saída para nós numéricos. Este padrão de ligação pode ser interpretado como as dimensões numéricas de um conceito.

Observando a Figura 21, mostrada anteriormente, os termos **mede** e **medindo** são ligados a três valores numéricos. Esta marcação pode ser interpretada pelo modelo como sendo uma característica tridimensional. Também, o termo **Volume** é ligado a somente um número, o que indica uma característica unidimensional. A mesma abordagem pode ser adotada para identificação de características bidimensionais. A frase a seguir denota um cisto bidimensional medindo 11 por 13 milímetros:

"Cisto medindo 11 x 13 mm."

Ligando o termo **Cisto** ao termo **medindo**, e o termo **medindo** aos dois valores numéricos, conforme a Figura 26, interpreta-se que os números **11** e **13** estão semanticamente ligados ao termo **medindo**. Dessa forma, **11** e **13** seriam as dimensões do conceito 'medindo' ligado ao conceito 'cisto'.

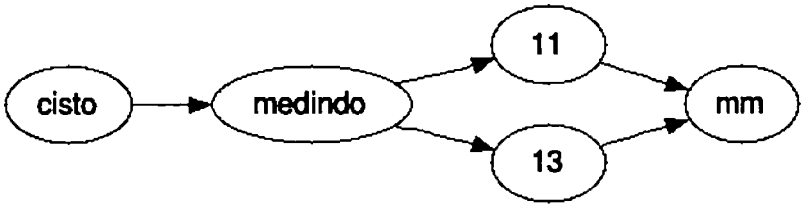

Figura 26: Ligação de dois tokens numéricos indicando um conceito bidimensional.

A identificação de valores multidimensionais pode ser alcançada sem uso de recursos semânticos ou linguísticos adicionais. Para o modelo, será importante adicionar na aresta uma característica que indica qual é a dimensão máxima para um determinado conceito, ou seja, com quantos nós numéricos um certo conceito poderá se relacionar ao mesmo tempo. Uma forma genérica de indicar esta característica para todas as arestas, e não somente para as arestas que atinjam nós numéricos, é registrar a cardinalidade máxima da ligação entre dois conceitos quaisquer. Dessa forma, a identificação de valores multidimensionais pode usar esta característica para determinar as dimensões de um conceito. A Figura 27 mostra parte do grafo da Figura 25 com a característica da cardinalidade máxima 'n' adicionada às arestas. Segundo a interação do usuário especialista, o conceito 'medir' será no máximo tridimensional. Já o conceito 'volume' é no máximo unidimensional.

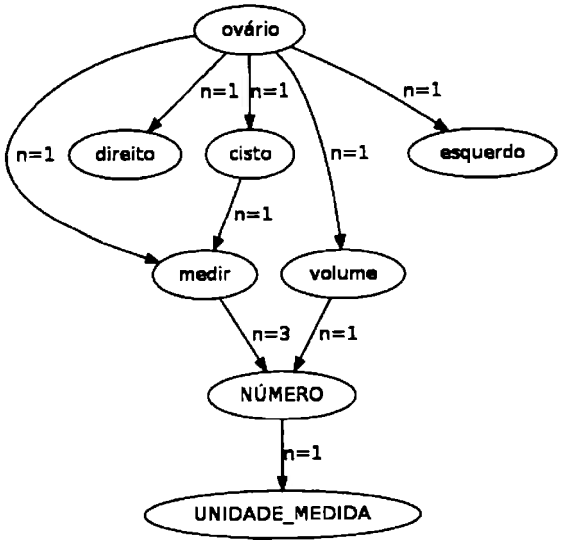

Figura 27: Grafo com os nós numéricos e de unidade de medidas substituídos pelos seus conceitos.

5.5.6 Identificação de intervalos

Dependendo do método utilizado para capturar os dados da fisiologia ou da patologia do paciente, as medidas podem ser imprecisas e algumas características são expressas usando valores relativos, ou intervalos, e não valores absolutos. Exemplo:

"Cisto com volume entre 11 e 13 cm³."

Essa é uma característica interessante a ser acoplada ao modelo, pois pode ser identificada a partir da ligação entre dois elementos numéricos, como mostra a Figura 28:

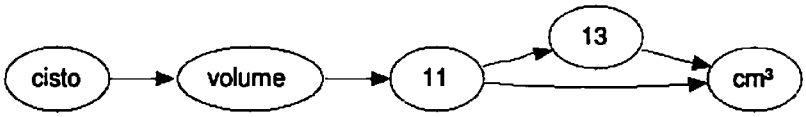

Figura 28: Ligação de dois elementos numéricos indicando um intervalo.

É importante diferenciar a marcação de valores de intervalo dos valores multidimensionais, pois na figura acima o número **11** está ligado ao **13**. Já se o termo **volume** estivesse ligado aos dois números, o método poderia interpretar que ambos valores indicam uma bidimensionalidade de **volume**. A ligação entre dois números terá a mesma semântica da preposição 'entre' e da conjunção 'e'.

Vale ressaltar que ao detectar a ligação entre dois números, o modelo poderá analisar os termos circunvizinhos, independentemente da linguagem utilizada no conteúdo, e identificar a preposição e conjunção da língua atualmente utilizada. Esta informação poderá servir para o método em futuras interações.

5.5.7 Representação de sentença de negação

A extração de informações estruturadas de um documento não é uma tarefa trivial, como tem sido discutido até aqui. Ainda mais quando os documentos apresentam sentenças de negação. No trecho do laudo a seguir é mostrado um exemplo desse desafio:

" Fundo de saco: Ausência de coleções anormais..."

A sentença acima descreve que numa região denominada *fundo de saco* não foram observadas coleções anormais. A expressão 'coleções anormais' indica um problema, uma situação patológica. Se a expressão 'fundo de saco' for ligada diretamente à expressão 'coleções anormais' haverá perda de uma informação importante, a informação fornecida pelo termo **Ausência**, que nega a presença de uma patologia, ou seja, indica que tudo está normal.

O modelo proposto deverá ser capaz de lidar com sentenças de negação para possibilitar uma extração coerente de informações dos laudos. Chapman (et al. 2001) propõe uma abordagem baseada em expressões regulares que identificam a presença de termos de negação em uma sentença e anota a sentença como negativa. É uma abordagem simples e que apresenta alto grau de dependência da linguagem e do domínio. O livro de Dowty (1994) reúne uma coleção de artigos de importantes autores da área de análise semântica utilizando recursos mais formais da lógica clássica para tratamento de itens de negação. Resta definir, com o andamento da pesquisa, quais técnicas poderão ser acopladas ao modelo para que sentenças negativas sejam automaticamente reconhecidas.

Analisando palavras como **não** e **sem** é possível verificar característica que as sucedem para identificar a negação. Como por exemplo, na Figura 25, mostrada na página 54, o conceito 'ovário' está ligado ao conceito 'não visualizado' e o conceito 'útero' está ligado ao conceito 'visualizado'. É possível inferir, conhecendo a semântica do termo **não**, que 'não visualizado' é a negação do conceito 'visualizado'.

5.5.8 Compartilhamento de valores entre várias características

Ao descrever o que é observado em uma imagem, o especialista pode compor uma sentença com várias características de um mesmo órgão. A descrição pode ser detalhada ou mais resumida. Quando detalhada, o método de ligação entre os conceitos poderá inferir facilmente as relações, como por exemplo, no trecho abaixo o termo **contornos** poderá ser ligado ao termo **regulares**, e o termo **ecotextura** poderá ser ligado ao termo **heterogênea**:

"Útero... os seus contornos são regulares e a ecotextura é heterogênea."

No caso de uma sentença resumida, que é comum em laudos médicos, um mesmo valor está relacionado à mais de uma característica:

"Útero... contorno e textura normais"

Na sentença acima, o termo **normais** é referente aos termos **contorno** e **textura**. No entanto, submetendo esta sentença a um analisador gramatical[13], o adjetivo **normais** aparece relacionado somente com **textura**. Neste caso o método deverá identificar que o conceito 'normais' se refere tanto ao conceito 'contorno', quanto ao conceito 'textura'. Esta identificação será possível se for considerado que alguns nós no grafo conceitual são folhas e outros são intermediários. Os nós intermediários refletem conceitos que, na maioria das vezes, estão ligados a outros conceitos. Dessa forma, os nós folhas tendem a ser possíveis valores de características definidas nos nós intermediários.

A Figura 29 mostra um grafo conceitual extraído das sentenças acima a partir da ligação dos conceitos principais. É possível verificar a existência de nós folhas e de nós intermediários. O conceito 'contorno' possui duas arestas de saída para os conceitos 'regular' e 'normal', que são nós folhas. Dessa forma, pode ser interpretado que regular e normal são possíveis valores para contorno. Logo, se o conceito contorno aparecer em um laudo, o método poderá buscar por possíveis valores referenciados mais a frente e tentar estabelecer uma ligação. O mesmo acontece com o conceito 'ecotextura', e com os conceitos 'transvaginal', 'direito', 'esquerdo', 'vazia', 'visualizado' e 'não visualizado' mostrados na Figura 25.

[13] http://comunidade.cogroo.org/index.html

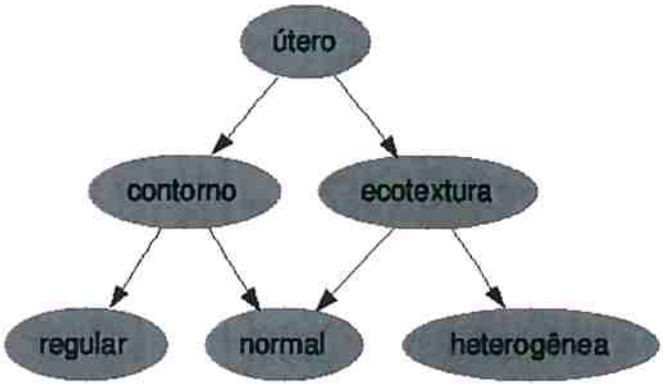

Figura 29: Grafo conceitual mostrando nós intermediários e nós folhas.

Durante o desenvolvimento do projeto descrito no **Capítulo 4** foi identificada uma lista de possíveis nós folhas. Esses nós representam os valores das características de um órgão, e geralmente não possuirão arestas de saída durante o processo de ligação dos conceitos. A Tabela 8 mostra a lista dos candidatos a nós folhas e a Erro: Origem da referência não encontrada mostra a lista de candidatos a nós intermediários, ou seja, que denotam características dos órgãos.

Tabela 8: Lista com termos que representam valores de características observadas em exames de Ultrassonografia Pélvica e Transvaginal.

anormal	grosso	pequeno
aumentado	habitual	pós-menopausada
ausência	heterogêneo	presença
ausente	homogêneo	regular
avf	inicial	separado
bocelados	interposição	simples
complexo	livre	sólido
convexo	médio	típico
evidente	mvf	tópico
fino	normal	vazio
grande	normoposicionado	virtual
grosseiro	padrão	

Tabela 9: Lista com termos que representam características observadas em exames de Ultrassonografia Pélvica e Transvaginal.

contorno	forma	repleção
dimensão	limite	retroversão
ecotextura	massa	textura
eixo	medir	vascularização
espessura	parede	volume

5.5.9 Relação do conceito com seu contexto

Ciravegna (2001) discute que a identificação de um conceito dentro de um documento com somente um assunto é muito mais fácil do que em documentos em que são tratados diversos assuntos. Nos laudos médicos são anotados os aspectos de vários órgãos observados durante a realização do exame. Logo, nos laudos são encontrados diversos assuntos. Assim, ao identificar o conceito 'volume', por exemplo, é necessário identificar qual órgão ou patologia ele referencia. No trecho do laudo a seguir é possível verificar a ocorrência do conceito 'medindo' duas vezes.

> ***... Ovário direito: Medindo 3,1 x 2,2 x 2,3 cm nos seus maiores eixos. Volume de 3,4 cm³. Aprsentando uma imagem cistica, de aspecto simples, medindo 21 mm...***

Como discutido anteriormente, os termos linguísticos denotam o contexto em que os conceitos estão inseridos e a ligação entre os mesmos. No entanto, os laudos apresentam uma incompletude linguística e a atual proposta pretende dispensar ao máximo o uso de fontes semânticas. Assim, o modelo deve identificar as relações entre os conceitos sem usar, por exemplo, a árvore gramatical das sentenças. A única informação que ele possui é o grafo conceitual criado pelo especialista. A Figura 30 apresenta um grafo conceitual parcial para exemplificar como o contexto pode ser identificado a partir do grafo.

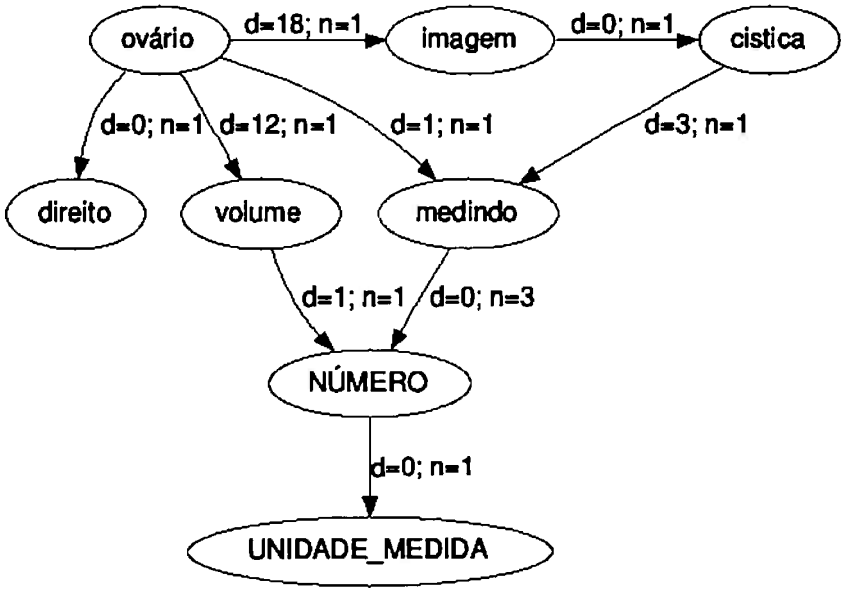

Figura 30: Grafo conceitual extraído a partir do trecho do laudo.

Tanto o conceito 'ovário' quanto o conceito 'cistica' estão ligados ao conceito 'medindo'. Ao analisar o trecho do laudo, utilizando uma abordagem da esquerda para a direita, o método inicia a análise pelo termo **ovário** e associa ele ao conceito 'ovário' do grafo. Seguindo a análise, é encontrado o termo **direito** que é associado ao conceito 'direito' do grafo. Então, as arestas de entrada no conceito 'direito' são analisadas para verificar se algum conceito associado a ele já foi encontrado. Neste caso, o conceito 'ovário' foi identificado recentemente e é origem de uma aresta de entrada do conceito 'direito', neste momento os metadados das arestas são analisados para encontrar o valor da função grauDeConfiança().

Logo, a relação de 'ovário' e 'direito' é identificada no documento. O próximo termo a ser analisado é **Medindo** que é associado ao conceito 'medindo'. Analisando suas arestas de entrada no grafo, dois conceitos relacionados são identificados, o conceito 'ovário' e 'cistica'. Nenhuma referência recente do conceito 'cistica' foi encontrada e sim do conceito 'ovário'. Logo, a relação de 'ovário' e 'medindo' é identificada. Assim, segue a análise do documento com a identificação e relação dos conceitos, de acordo com a hierarquia apresentada no grafo conceitual e sua ordem de aparição no documento. A Figura 32 mostra o grafo R que deve ser gerado a partir da análise do trecho do laudo mostrado anteriormente.

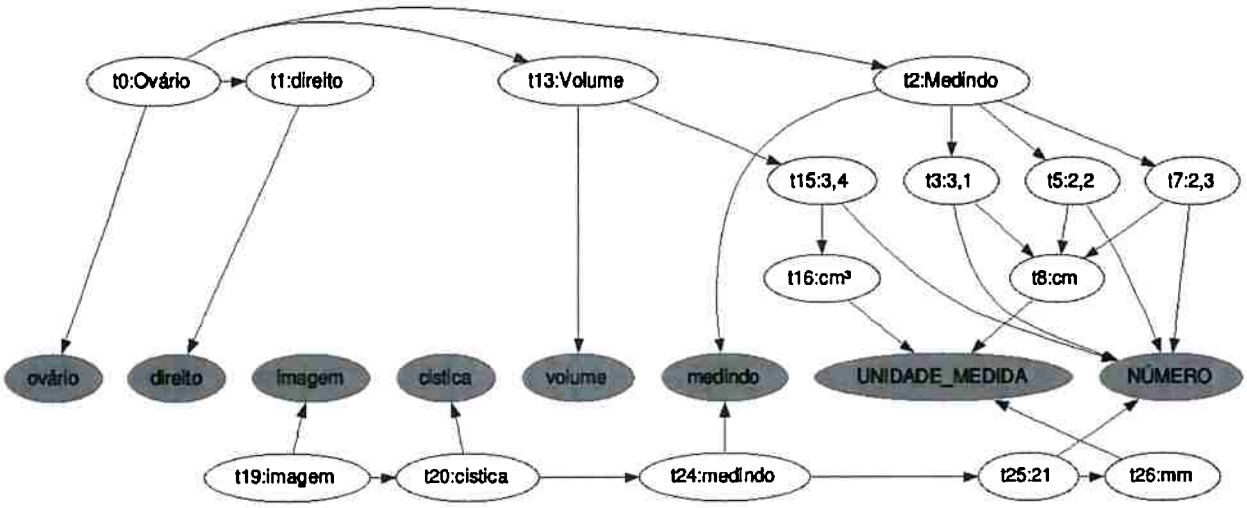

Figura 31: Grafo R resultante da análise do trecho do laudo segundo o grafo conceitual da Figura 30.

Analisando os laudos médicos, uma característica de um órgão é sempre anotada após a citação do órgão, nunca adiante, ou seja, as sentenças estão na forma direta, sempre o sujeito vem antes do predicado. Isto é importante para que o método identifique os conceitos anteriormente citados para estabelecer as ligações. No entanto, deverá ser verificado como o método poderá lidar com sentenças na forma indireta, ou seja, onde o predicado aparece antes do sujeito. Uma possibilidade é analisar se o usuário ligou um termo ao seu vizinho antecedente, ou seja, a direção da ligação é inversa ao fluxo original do texto. Assim, o método deverá anotar que este conceito pode ser citado antes do seu sujeito. Dessa forma, o método não poderá usar somente uma abordagem de cima para baixo na análise do laudo.

A próxima seção analisa como as informações extraídas poderão ser processadas. E as seções seguintes apresentam o protótipo desenvolvido e como o modelo deverá ser testado e validado.

5.6 Armazenamento e busca das informações

O modelo proposto deverá processar as informações desestruturadas dos laudos e gerar grafos contendo as informações estruturadas sobre cada laudo. Esses grafos podem compor uma base de conhecimento e serem armazenados usando o formado RDF. O formato OWL foi avaliado, porém, para a simplicidade semântica do modelo, o formato RDF mostrou-se mais adequado, além de contar com um

conjunto bastante abrangente de ferramentas. Usando o RDF, cada aresta do grafo gerará uma tripla (sujeito, predicado, objeto). O sujeito será o nó origem da aresta. O predicado indicará o tipo de ligação que os nós possuem. E o objeto será o nó destino da aresta. Inicialmente, o modelo possui dois tipos de ligações: uma para ligar os termos do laudo com os conceitos do modelo, arestas 'é um', e outro para ligar os termos entre si, arestas 'tem um'. A Figura 32 mostra um exemplo de grafo com os tipos das arestas.

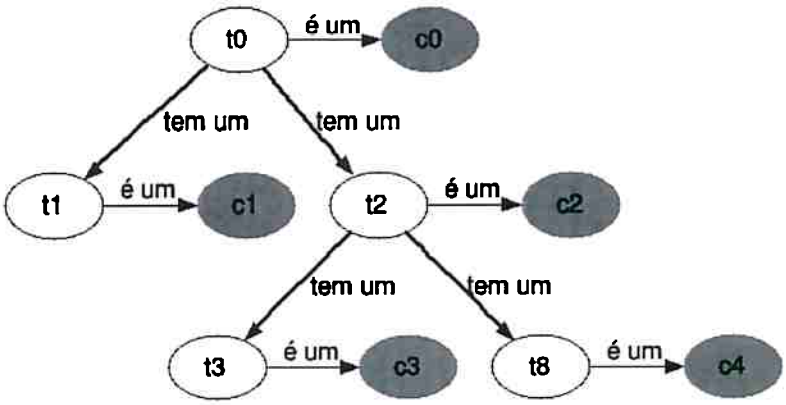

Figura 32: Grafo que representa o conhecimento armazenado em um laudo usando as arestas 'é um' e 'tem um'.

Outro objetivo do modelo proposto é possibilitar a busca pelos laudos usando uma frase em linguagem natural, bastante semelhante a utilizada durante a composição dos laudos. Assim, o especialista médico não precisa aprender uma linguagem técnica de busca de conhecimento, como SparQL.

Inicialmente, será utilizada a mesma técnica descrita na **Seção 4.4**, e inspirada no trabalho de Montes-y-Gómez (2000), que consiste em analisar a frase de busca utilizando o mesmo grafo conceitual aplicado na extração das informações. Assim, os conceitos referenciados na frase são extraídos e um grafo é gerado. Esse grafo pode ser traduzido para SparQL e submetido ao mecanismo de busca em RDF.

5.7 Protótipo do método

O atual projeto propõe uma modelagem visual e intuitiva para anotação de conceitos em conteúdos textuais, gerando um modelo em grafo conceitual, sem a necessidade de escrita de expressões de extração ou outro conhecimento mais técnico de NLP. Para iniciar os testes, foi construído um protótipo que permite digitar o conteúdo do laudo, realizar alguns pré-processamentos e permitir a ligação entre os termos do laudo. A Figura 33 mostra uma versão do protótipo sendo executada em um navegador de páginas HTML.

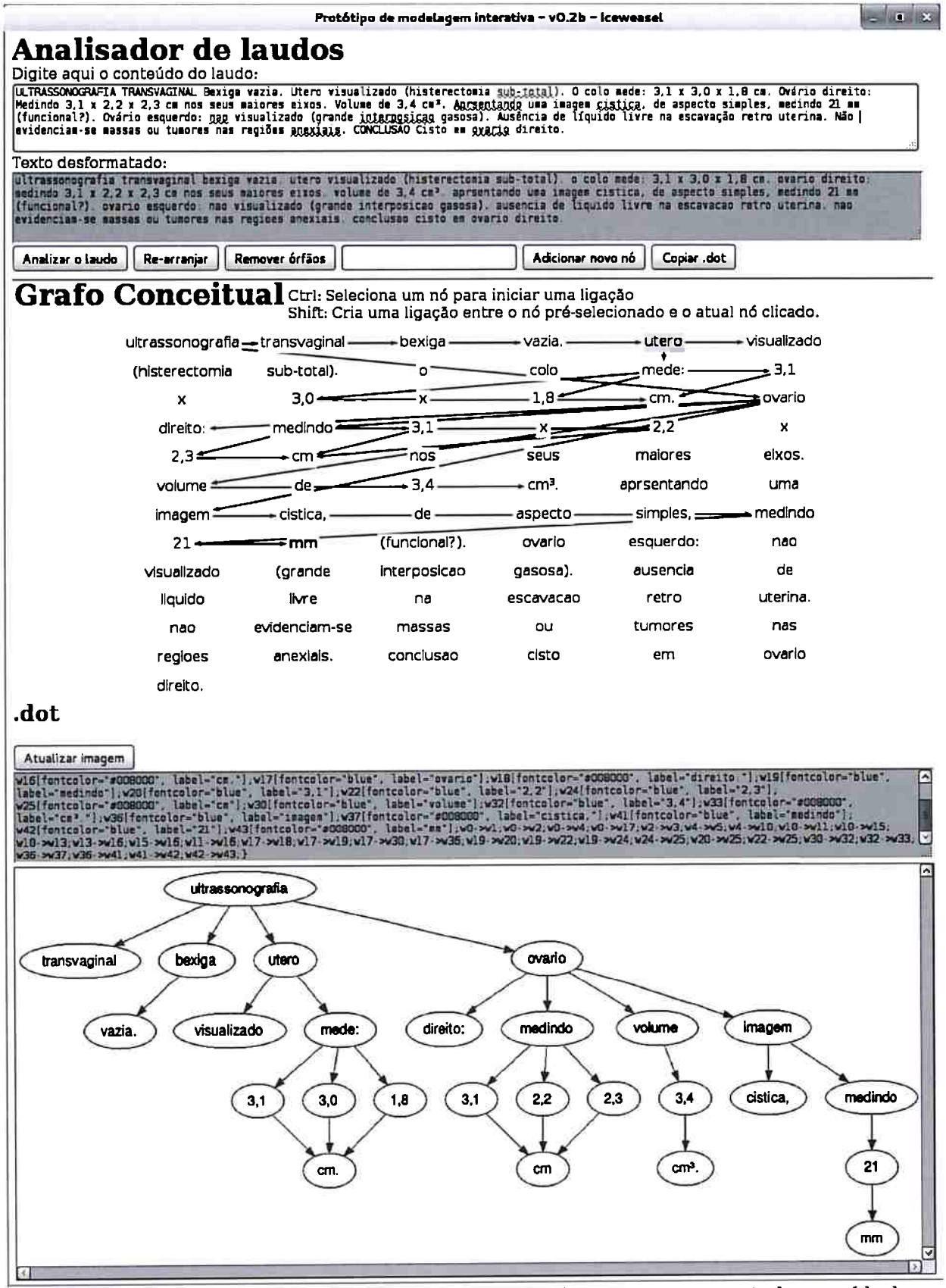

Figura 33: Protótipo construído para testar a criação de grafos conceituais a partir do conteúdo de um laudo.

O especialista deve estabelecer ligações entre as palavras clicando sobre elas. Para definir a origem de uma ligação deve-se clicar sobre uma palavra com a tecla Ctrl pressionada. Clicando sobre outra palavra com a tecla Shift pressionada, uma ligação entre a palavra origem e a atual palavra é criada. Cada termo pode ser arrastado para uma melhor organização visual do conteúdo que está sendo analisado.

5.8 Avaliação e validação do modelo

Os resultados apresentados no **Capítulo 4** são de um modelo bastante específico e limitado, mas que já mostra sua utilidade. A atual proposta apresenta um modelo mais genérico e formal. Para efetuar sua validação será utilizado um conjunto limitado de laudos manualmente anotados. Também poderá ser verificada a possibilidade da aplicação do método a corpus com anotação semântica, como o PMBOK (2001). No entanto, isso só será possível se os resultados alcançados conseguirem uma independência linguística, uma vez que o PMBOK e outro corpus são para o Inglês.

Como descrito nas fases 1, 2 e 3, alguns conceitos serão ligados em um pequeno número de laudos. O método construirá o modelo conceitual e será verificada a precisão e a cobertura dos resultados alcançados pelo modelo de acordo com o conjunto de teste.

Também pretende-se avaliar a eficiência do método, comparando seu desempenho no processo de indexação com os modelos de índices descritos no **Capítulo 4**, o índice invertido convencional e o índice ontológico. Esta avaliação mostrará se o processo de aplicação do modelo na extração consegue indexar o mesmo número de laudos em maior ou menor tempo que os demais.

Outra parte a ser avaliada é a facilidade para especificação e extração dos conceitos usando o método proposto. Deverá ser verificada a capacidade de um especialista médico iniciar a marcação de conceitos após algumas breves orientações. Essa avaliação é mais subjetiva e poderá ser realizada através de testes práticos e aplicação de questionários. A intenção desta avaliação é mostrar que o método poderá ser facilmente usado por pessoas com praticamente nenhum conhecimento em PLN. Esta característica do método proposto é importante, uma vez que praticamente todas os trabalhos analisados apresentam um certo grau de complexidade para a configuração e inicialização das regras de extrações, dependendo muitas vezes da definição de expressões que seguem gramáticas próprias.

Referências Bibliográficas

Aggarwal, Charu C., and ChengXiang Zhai. *Mining text data.* Springer, 2012.

Aires, Rachel Virgínia Xavier. *Implementação, adaptação, combinação e avaliação de etiquetadores para o português do brasil.* Dissertação de Mestrado, ICMC-USP, São Carlos - SP, 2000.

Alpaydin, Ethem. *Introduction to Machine Learning.* MIT Press, p. 9, ISBN 978-0-262-01243-0, 2010..

Barbosa, Flavio. *Metodologia para estruturação de informações de laudos radiológicos.* Tese Faculdade De Medicina De Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2013.

Bechhofer, Sean, et al. *OWL web ontology language reference,* 2004. http://www. w3. org/TR/owl-ref. Acessado em Agosto 2014 .

Black, William J., et al. *CAFETIERE: Conceptual Annotations for Facts, Events, Terms, Individual Entities. and RElations.* Technical Report TR–U4. 3.1, Department of Computation, UMIST, Manchester, 2005.

Bodenreider, Olivier. *The unified medical language system (UMLS): integrating biomedical terminology.* Nucleic acids research 32.suppl 1 (2004): D267-D270.

Brascher, Marisa. *A ambiqüidade na recuperação da informação.* Artigo 05. DataGramaZero - Revista de Ciência da Informação, v.3 n.1, ISSN1517-3801, 2002.

Brasil. *Livro Branco: Ciência, Tecnologia e Inovação.* Ministério da Ciência e Tecnologia, Brasília, ISBN85-88063-04-2, 2002.

Breitman, Karin, Marco Antonio Casanova, Walt Truszkowski. *Semantic Web: Concepts, Technologies and Applications: Concepts, Technologies and Applications.* Springer, 2007.

Carvalheira, Luiz Carlos da Cruz. *Método semi-automático de construção de ontologias parciais de domínio com base em textos.* Dissertação Escola Politécnica da Universidade de São Paulo. Universidade de São Paulo, 2007.

Chapman, Wendy W., et al. *A simple algorithm for identifying negated findings and diseases in discharge summaries.* Journal of biomedical informatics 34.5, 2001: 301-310.

Ciravegna, Fabio. *Adaptive information extraction from text by rule induction and generalisation.* International Joint Conference on Artificial Intelligence. Vol. 17. No. 1. Lawrence Erlbaum Associates Ltd, 2001.

Cunningham, Hamish. *GATE, a general architecture for text engineering.* Computers and the Humanities 36.2, 2002: 223-254.

DATASUS-CNES, *Ministério da Saúde - CNES – Recursos Físicos – Equipamentos – Brasil - SIAB. Informações Estatísticas. Produção e Marcadores.*

http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/equipobr.def. Acessado em Julho de 2016.

DATASUS-SIAB, *Ministério da Saúde - Sistema de Informação de Atenção Básica - SIAB. Informações Estatísticas. Produção e Marcadores.* http://tabnet.datasus.gov.br/cgi/deftohtm.exe?siab/cnv/SIABPBR.DEF. Acessado em Julho de 2016.

Dias-da-Silva, Bento C., Ariani Di Felippo, and Ricardo Hasegawa. *Methods and tools for encoding the wordnet.br sentences, concept glosses, and conceptual-semantic relations.* Computational Processing of the Portuguese Language. Springer Berlin Heidelberg, 2006. 120-130.

Ding, Jing, et al. *Extracting biochemical interactions from MEDLINE using a link grammar parser.* Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on. IEEE, 2003.

Dowty, David. *The role of negative polarity and concord marking in natural language reasoning.* Proceedings of SALT. Vol. 4. 1994.

Ciravegna, Fabio, and Wilks Yorick. *Designing Adaptive Information Extraction for the Semantic Web in Amilcare.* In S. Handschuh and S. Staab, editors, Annotation for the Semantic Web. IOS Press, Amsterdam, 2003.

Feldman, Ronen, and James Sanger, eds. *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge University Press, 2007.

Fellbaum, Christiane. *WordNet.* Blackwell Publishing Ltd, 1998.

Freitas, Larissa Astrogildo de, e Renata Vieira. *Ontologias e língua portuguesa.* Anais do CELSUL 1.2, 2008.

Goh, Cheng Hian. *Representing and reasoning about semantic conflicts in heterogeneous information systems.* Dissertation Massachusetts Institute of Technology, 1996.

Haythornthwaite, Caroline, and Anatoliy Gruzd. *A noun phrase analysis tool for mining online community conversations.* Communities and Technologies 2007. Springer London, 2007. 67-86.

Hearst, Marti A. *Automated Discovery of WordNet Relations,* in WordNet: An Electronic Lexical Database, Christiane Fellbaum (ed.), MIT Press, 1998.

Heinze, Daniel T., et al. *Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology.* Journal of the American Medical Informatics Association 15.1, 2008: 40-43.

Heinze, Daniel T., Mark L. Morsch, and John Holbrook. *Mining free-text medical records.* Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001a.

Heinze, Daniel T., et al. *LifeCode: A deployed application for automated medical coding. Ai Magazine* 22.2, 2001b: 76.

Honorato, Daniel De Faveri. *Metodologia de transformação de laudos médicos não estruturados e estruturados em uma representação atributo-valor*. Dissertação, ICMC-USP, 2008.

Ijntema, Wouter, et al. *A lexico-semantic pattern language for learning ontology instances from text*. Web Semantics: Science, Services and Agents on the World Wide Web 15, 2012: 37-50.

BIPM. *International Bureau of Weights and Measures, The International System of Units (SI)*, 8th edition, ISBN 92-822-2213-6, BIPM. 2006.

Jacobs, Paul S., George R. Krupka, and Lisa F. Rau. *Lexico-Semantic Pattern Matching as a Companion to Parsing in Text Understanding*. HLT. 1991.

Kruskal, Joseph B. *An overview of sequence comparison: Time warps, string edits, and macromolecules*. SIAM review 25.2, 1983: 201-237.

Lapa, Remi Correia; Renato Fernandes Corrêa. *O estado da arte da pesquisa sobre indexação automática realizada no Brasil no âmbito da ciência da informação (1973-2012)*. Encontro Nacional De Pesquisa Em Ciência Da Informação, 14., UFSC, Florianópolis, 2013.

Lipscomb, Carolyn E. *Medical subject headings (MeSH)*. Bulletin of the Medical Library Association 88.3, 2000: 265.

Madhyastha, Harsha V., N. Balakrishnan, and K. R. Ramakrishnan. *Event information extraction using link grammar*. Research Issues in Data Engineering: Multi-lingual Information Management, 2003. RIDE-MLIM 2003. Proceedings. 13th International Workshop on. IEEE, 2003.

Maynard, Diana, Yaoyong Li, and Wim Peters. *Nlp techniques for term extraction and ontology population*. Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. 2008.

Montes-y-Gómez, M., López-López, A. and Gelbukh, A. "Information Retrieval with Conceptual Graph Matching", Lecture Notes in Computer Science, Vol. 1873, Springer Verlag, 2000, pp. 312-321.

Mukherjea, Sougata, and Bhuvan Bamba. *BioPatentMiner: an information retrieval system for biomedical patents*. Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. VLDB Endowment, 2004.

Narciso, Everton N., Fátima LS Nunes, e Márcio E. Delamaro. *Seleção de Casos de Teste Utilizando Conceitos de Variabilidade: Uma Revisão Sistemática*. Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo–SP, 2011: 115-125.

Netto, Oscar Picchi, et al. *Uma Metodologia para Estruturação de Laudos Médicos usando Ontologias*. XI Workshop de Informática Médica, XXXI CSBC, Natal 2011: 29.

Oleynik, Michel, et al. *Performance analysis of a POS tagger applied to discharge summaries in Portuguese*. Studies in health technology and informatics 160.Pt 2, 2009: 959-963.

Oleynik, Michel. *Extração de informações de narrativas clínicas*. Dissertação. Instituto de Matemática e Estatística, Universidade de São Paulo, 2013.

Oliveira, Lutero M. de. *Radiologia e Diagnóstico por Imagem - Ética, Normas, Direitos e Deveres dos Médicos Imaginologistas*, Colégio Brasileiro de Radiologia e Diagnóstico por Imagem – CBR, 2012.

Pellizzon, Rosely de Fátima. *Pesquisa na área da saúde: 1. Base de dados DeCS (Descritores em Ciências da Saúde).* Acta cir. Bras 19.2, 2004: 153-163.

PMBOK. *A Guide to the Project Management Body Of Knowledge (PMBOK® Guide).* 4th Edition, Project Management Institute. ISBN: 978-1-933890-51-7, 2001.

Popov, Borislav, et al. *KIM-a semantic platform for information extraction and retrieval*. Natural language engineering 10.3-4, 2004: 375-392.

Serapião, Paulo Roberto Barbosa, Kátia Mitiko Firmino Suzuki, e P. M. Azevedo-Marques. *Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia.* Radiol Bras 43, 2010: 103-7.

Soderland, Stephen. *Learning information extraction rules for semi-structured and free text*. Machine learning 34.1-3, 1999: 233-272.

Sowa, John F. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., 1984.

Ullmann, Stephen. *Semantics: an introduction to the science of meaning*. Barnes & Noble, 1962.

Zhou, Xiaohua, et al. *Approaches to text mining for clinical medical records*. Proceedings of the 2006 ACM symposium on Applied computing. ACM, 2006.

Zhou, Xiaohua, et al. *Converting semi-structured clinical medical records into information and knowledge*.Data Engineering Workshops, 2005. 21st International Conference on. IEEE, 2005.