ONTOLOGIAS PARA REPRESENTAÇÃO DE CONHECIMENTO EXTRAÍDO DE LAUDOS DE RADIOLOGIA TORÁCICA.

Leandro Zerbinatti[1], Lincoln de Assis Moura Jr[2]

Abstract — *This work comprises the design and evaluation of a method for the extraction of knowledge from free-text chest X-ray reports. Our final goal is to be able to automatically transform the free-text reports into ontology using OWL with Protégé documents conveying relevant information. To achieve that a set of 2000 reports, written in Portuguese, formed the basis from which a corpus was derived. A classical approach was taken to analyzing the data: tokens were found and n-grams were identified, their frequency and relevance computed. All relevant 160 n-grams were mapped into SNOMED, LOINC and/or DICOM classes and into an HL7/CDA template. This approach has made it possible to convert the free-text reports into CDA documents.*

Index Terms — *Health Information Standards, Health Information Ontology, Health Information Systems, Natural Language Processing.*

INTRODUÇÃO

A Biomédica é um dos campos do conhecimento mais recentes da Engenharia. Desta área inovadora surge um vasto campo na maneira de formar especialistas e como aplicar esse conhecimento.

A Engenharia Biomédica conjuga conhecimentos das ciências exatas e ciência da saúde, além da aplicação de conhecimentos de química, física, biologia, medicina, e técnicas de engenharia química, engenharia eletrônica, engenharia biológica, engenharia de materiais, entre outras.

Formar engenheiros de forma que eles adquiram o conhecimento de tantas áreas do conhecimento e com tantas informações geradas e inovadas é tarefa hercúlea.

Uma forma de minimizar esse esforço é deixar disponíveis os conhecimentos da área de saúde de maneira formal, utilizando padrões de representação de informação na saúde, garantindo o reuso dos dados e a interoperabilidade entre os sistemas.

Assim engenheiros e técnicos poderiam se concentrar nas soluções inovadoras dessa rica área, utilizando o conhecimento consensual médico.

Alguns estudos já discorreram sobre este problema, porém estão concentrados na língua inglesa não derivando todos esses conhecimentos para a língua portuguesa que é a 7ª língua mais falada no mundo e a mais falada no sul do equador.

Nosso objetivo é extrair o conhecimento médico a partir de relatórios de radiologia do tórax, escrito em português, utilizando técnicas de processamento estatístico de linguagem natural, a identificação do conteúdo semântico desses relatórios, a organização de conteúdos de acordo com uma ontologia existente (SNOMED-CT) e a criação de relatórios estruturados em conformidade com a norma HL7 CDA (*Clinical Document Architecture*) e, portanto, compatível com o DICOM e LOINC.

CONHECIMENTO E REPRESENTAÇÃO DO CONHECIMENTO

O conhecimento médico está diretamente ligado a aspectos do conhecimento experiencial, do conhecimento baseado em evidências e no conhecimento formal.

O conhecimento experiencial vem da vivência do profissional médico e ele pode ser sistematizado através de informações constantes nas bases de dados dos pacientes, envolvendo processos nos quais o médico reconhece um padrão clínico que ele já conhecia de experiências anteriores. O conhecimento baseado em evidências disponíveis são as evidências acumuladas ao longo do tempo e sumarizadas pelos processos estatísticos adequados. O conhecimento formal se baseia na literatura médica, periódicos, livros e outras formas de armazenamento incluindo ainda processos de dedução. [1]

Em todas essas formas de conhecimento a linguagem é a forma mais utilizada para representar o conhecimento.

Na representação da linguagem natural utilizamos símbolos alfa-numéricos compondo palavras e sentenças de forma a expressar, de forma livre, o conhecimento sobre determinada área de domínio.

Com a criação e uso do computador, outras formas de representação foram criadas para que sejam manipuladas e utilizadas a partir de computadores e a rede de informação que os interligam.

Processamento Estatístico de Linguagem Natural

Existem várias línguas faladas no mundo. Este fato decorre das explorações de terras, imigrações dos povos e relações comerciais entre os países.

[1] Escola Politécnica da Universidade de São Paulo, Fundação Aplicações de Tecnologias Críticas -Atech; leandro.zerbinatti@poli.usp.br
[2] Escola Politécnica da Universidade de São Paulo, ZILICS – Sistemas de Informação na Saúde; lincoln.moura@zilics.com.br

Em números absolutos, o inglês não é a língua mais falada no mundo, porém aquela mais utilizada comercialmente.

Em números de falantes nativos, o ranking das línguas segundo o livro "*The Etnologue languages of the world*" [2] é: 1º - Chinês (1.213.000,00), 2º - Espanhol (329.000.000), 3º - Inglês (328.000.000), 4º - Árabe (221.000.000), 5º - Hindi (182.000.000), 6º - Bengali (181.000.000), 7º - Português (178.000.000).

Segundo informações do *World Health Organization* em seu projeto denominado ePORTUGUESe, o português é utilizado por oito países em quatro continentes, conforme demonstrado na figura 1.

FIGURA. 1
PAÍSES QUE FALAM A LÍNGUA PORTUGUESA.

Independente da linguagem elas são representados por as palavras podem ser definidas como uma seqüência ordenada de símbolos de tamanho finito. Ao conjunto de símbolos chamamos de alfabeto. Em alguns casos o alfabeto utilizado representa literalmente o alfabeto (a, b, c, d, e...z) em outras circunstância o alfabeto também é composto de outros caracteres como números e símbolos (0,1,2,3...9, +, *, !, ?).[3]

Representamos o alfabeto grego com o símbolo Σ, e que todas as palavras formadas por uma seqüência de símbolos de Σ chamaremos de Σ+. O conjunto Σ é finito, já o conjunto Σ+ é infinito. [3]

Para incluir todos os caracteres, incluindo os símbolos, por exemplo, o espaço, utilizaremos o símbolo ε, que denota um caractere sem símbolo. Esta falta de símbolo não é um elemento de Σ ou de Σ+, mas um novo conjunto ε. A equação (1) define Σ* como a união entre o conjunto Σ+ e o conjunto ε.

$$\Sigma^{*} = \Sigma^{+} \cup \{\varepsilon\} \qquad (1)$$

Portanto Σ* pode ser definido como o conjunto de palavras compostas por símbolos de tamanho zero (espaço) ou maior.

Para a formulação de uma semântica, não é necessário o uso de todas as palavras constantes em Σ*, mas somente aquelas que têm senso em um contexto particular, e isto têm relação com o significado das palavras. [3]

Com o uso de estatística no processamento de linguagem natural identifica-se que as palavras ocorrem em uma distribuição desigual. Algumas palavras são muito comuns e normalmente, as 100 palavras mais comuns ocorrem um pouco mais da metade das palavras que ocorrem em um corpus. [4]

Em 1949, George K. Zipf, em seu livro *"Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology"* descreveu uma material sobre o princípio do mínimo esforço, que discrimina a essência da condição humana.

A teoria de Zipf nos informa que se nós contarmos a freqüência de um tipo de palavra que ocorre na linguagem escrita de um corpus, e então listar as palavras em ordem de freqüência de ocorrência, na equação (2) exploramos a relação entre a freqüência de uma palavra f e a sua posição na lista chamada rank r.

$$k = f.r \qquad (2)$$

De acordo com a teoria de Zipf, tanto o indivíduo que está falando, quanto o indivíduo que está ouvindo, estão tentando minimizar o esforço.

Sobre o número de significados das palavras, Zipf argumenta que para o emissor conservar seu esforço, ele prefere utilizar somente uma palavra que contenha todos os significados, enquanto para o esforço do ouvinte ele preferirá que cada significado seja expresso com uma palavra diferente. Assumindo que estas forças são igualmente fortes, identifica-se que quanto maior o *rank* de uma palavra, maior o número de significados que ela assume. [4]

Ontologias

O termo original da palavra ontologia é a palavra aristotélica "categoria" que pode ser utilizada para classificar algo. Aristóteles apresentava categorias para classificar entidades e distinguir espécies diferentes do mesmo gênero. [5]

Atualmente a palavra ontologia é composta por outras duas "onto" e "logia", que significa o estudo ou o conhecimento do Ser, dos entes ou das coisas tais como são em si mesmas, real e verdadeiramente, correspondendo ao que Aristóteles chamara de filosofia primeira, isto é, o estudo do Ser enquanto Ser. [6]

O termo Ontologia foi utilizado em 1613 por Rudolf Gockel e Jacob Lorhard, aparentemente de maneira independente.

O termo Ontologia ganhou expressão nas áreas de Inteligência Artificial, Computação Lingüística e Teoria de Banco de Dados, em áreas de pesquisa de Engenharia do Conhecimento, Análise Orientada a Objetos e Desenho de Sistemas baseados em Agentes. [7].

Para Borst (1997) "Uma ontologia é uma especificação formal e explicita de uma conceitualização compartilhada", onde formal significa legível para computadores, especificação explícita significa os conceitos, propriedades, relações, funções, restrições, axiomas, compartilhado quer dizer conhecimento consensual e conceitualização diz respeito a um modelo abstrato de algum fenômeno do mundo real. [8].

Uma ontologia deve conter (1) Classes, utilizado para modelar os conceitos do domínio organizados em taxonomias, (2) Atributos, representam as características dos conceitos, (3) Relações, modelam os tipos de associações e conceitos, (4) Instâncias, representam especializações de uma determinada classe, (5) Função, representam casos especiais de relações e (6) Axioma, modelas sentenças que são geralmente verdadeiras.

As ontologias escritas atualmente em OWL são uma boa forma de representar o conhecimento e torná-lo reusável.

Arquitetura de Documentos Clínicos

A arquitetura de documentos clínicos ou *Clinical Document Architecture* (CDA) é um padrão para organizar os documentos clínicos possibilitando a troca de informações entre sistemas utilizando arquivos eletrônicos de documentos médicos.

Alguns trabalhos estão abordando o assunto de aplicações de técnicas de processamento estatístico de linguagem natural e a apresentação da semântica juntamente com o CDA.

Em um dos trabalhos foi utilizado um detector de limite de orações, um analisador estatístico de linguagem natural e um etiquetador no CDA. O trabalho foi comparado com um as descrições realizadas por autores médicos. [9]

Atualmente nosso trabalho contém em sua análise um detector de limite de orações, a composição em n-gramas analisando menores unidades de conhecimento e um analisador estatístico de linguagem natural. Atualmente estamos trabalhando na composição de documentos aderentes ao CDA.

Em outro trabalho [10] foi aplicado um módulo denominado de NLU (*Natural Language Understanding*) para fazer a análise sintática de diversos tipos de laudos, inclusive de radiologia torácica e a geração de arquivos XML compatíveis com o CDA.

Em um terceiro trabalho de referência foi descrito as lições aprendidas na tradução de conceitos complexos da área clínica e biomédica, escritos em linguagem natural, em formatos legíveis para o computador como CDA e padrões como o LOINC e o SNOMED-CT [11], não foi utilizado técnicas de processamento estatístico de linguagem natural.

MATERIAIS E MÉTODOS

Os dados apresentados neste trabalho se referem a 2000 laudos de radiologia torácica, digitados por médicos do Incor – Instituto do Coração de São Paulo, em um sistema designado para tal. O médico usuário do sistema pode selecionar sentenças e posteriormente editar estas sentenças.

Os laudos manipulados pelos autores não possuem nenhum tipo de dados de identificação.

Obviamente que os resultados se restringem a uma cultura local de doutores do Incor, porém não invalida o método utilizado.

A língua portuguesa é uma linguagem que utiliza muitos sinais diacríticos e nenhuma destas marcas foram digitadas pelos doutores. Esta ausência de sinais diacríticos provém das características do sistema utilizado para gerar os laudos. Como o conjunto de palavras é de um conjunto específico, não houve dificuldade em tratar essas palavras sem a correta grafia.

É importante salientar a forma de edição dos laudos de radiologia torácica adotada pelo Incor.

Na tela de descrição de laudos de raios-X de tórax o usuário pode selecionar frases a partir de uma lista de frases previamente descritas.

Esse processo inicia-se pelo digitador, que seleciona a frase e edita o texto liberando para o médico. O médico pode alterar a edição do texto até a liberação ao paciente.

Os métodos utilizados nesse trabalho podem ser descrito com o conjunto de passos abaixo:

1. *Tokenização* dos laudos.
2. Computação do número total de *tokens*.
3. Computação da freqüência das palavras.
4. Computação da constante de Zipf.
5. Obtenção do *n-gramas* (para n de 2 a 10).
6. Identificação de *n-gramas* que ocorrem mais de uma vez.
7. Deleção de *n-gramas* que contenham pontuações.
8. Deleção de *n-grams* que iniciem ou terminem com palavras de conexão como preposições e conjunções.
9. Mapeamento das palavras e os *n-gramas* com seu respectivo associado ao SNOMED-CT.
10. Construção de ontologias utilizando o Protegé.
11. Mapear as ontologias nas classes do HL7 CDA.

RESULTADOS

O Corpus

O corpus analisado nesse estudo foram dois mil (2000) laudos de radiologia torácica. Tipicamente esses laudos possuem características que descrevem; os tipos de equipamentos utilizados, os procedimentos realizados, os

achados e a conclusão. Nem todos os laudos apresentam essa tipificação. Abaixo um exemplo de laudo analisado.

tecnica:
exame realizado em equipamento de alta potencia, com tecnica de alta kilovoltagem, que fornece escala de contraste longa (baixo contraste).
laudo:
assimetria mamaria com irregularidades dos contornos.
osteopenia difusa.
artrose bilateralmente em ombros.
escoliose dorsal com achatamento de corpos vertebrais.
pincamento de cupula diafragmatica esquerda e opacidades imprecisas.
vascularizacao pulmonar normal.
hilos de configuracao, topografia e dimensoes normais.
imagem cardiaca normal.
aorta alongada.
arco medio normal.
espessamento cisural a esquerda.
condensacao de limites imprecisos em apice de hemitorax esquerdo (pneumonite actinica sequelar?).
conclusao:
reestadiamento de neoplasia mamaria.
ac.

Fica claramente identificado que as palavras não são acentuadas na origem da geração dos laudos.

Os Tokens e as Palavras

Após o processo de *tokenização* foram identificados 153.478 *tokens*. Confirmando os estudos de Zipf, a distribuição das 100 primeiras tem uma ocorrência elevada, e a maioria das palavras tem poucas ocorrências, conforme exemplificado na figura (2)

Listamos na tabela 1 o ranking dos dez *tokens* mais freqüentes nos experimentos. Esses *tokens* compõem 1067 tipos de palavras e ocorrências no corpus. Note que três dos dez itens apresentados no *rank* são pontuações, outros dois são preposição e conjunção, que não tem uma semântica associada à área medica e outros cinco são palavras com conceitos que podem ser associados ao conhecimento da área de saúde.

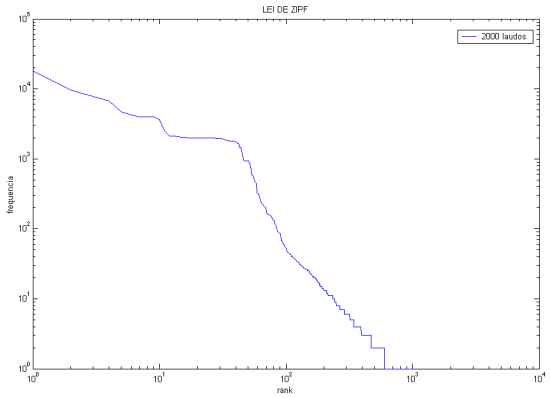

FIGURA. 2
A OCORRÊNCIA DE PALAVRAS NO CORPUS

TABELA I
RANK DOS TOKENS MAIS FREQUENTES

Rank	Tokens	Quantidade de Ocorrências
1	'.'	17.896
2	'de'	9.623
3	'normal'	7.832
4	','	6.750
5	'e'	4.743
6	':'	4.255
7	'tecnica'	4.006
8	'alta'	3.974
9	'contraste'	3.973
10	'pulmonar'	3.685

Os n-gramas

Foram obtidos os *n-gramas* com n= 2,3,4,5,6,7,8 e 9 e retirados os *n-gramas* que não se repetiam, por não oferecer representatividade estatística, além daqueles que na sua composição pontuações ou artigos, preposições ou conjunções no início ou final da sua formação.

O principal objetivo era identificar segmentos de conceitos representados através de *n-gramas* significativos para a área de saúde.

Na tabela II é demonstrado os *n-gramas* de maior ocorrência em cada n analisado.

TABELA II
RANK DOS TOKENS MAIS FREQUENTES

N	n-grama
N=2	'exame realizado.'
N=3	'tecnica de alta'
N=4	'escala de contraste longa'
N=5	'realizado em equipamento de alta,'
N=6	'exame realizado em equipamento de alta'
N=7	'exame realizado em equipamento de alta potencia'
N=8	'estruturas osseas e tecidos moles na parede toracica'
N=9	'coracao de dimensoes normais e morfologia compativel com o aumento'

Representação Ontológica e na Arquitetura de Documentos Clínicos

Após selecionar as palavras e os *n-gramas* com significado na língua portuguesa, esses foram comparados com o SNOMED – CT, que é uma documentação e relatórios sobre a terminologia clínica. Essa terminologia está associada a conceitos e relações entre as informações. Ele é dividido em uma hierarquia de classes.

Foi adotada esta hierarquia de classes para representar em ontologias, utilizando o Protégé.

Às classes, foram associadas às instâncias, que são as palavras e *n-gramas* que resultaram da análise do corpus e da verificação com o SNOMED-CT.

Nem todas as classes hierárquicas do SNOMED-CT foram utilizadas visto que as palavras e *n-gramas* cobriram

somente as classes de procedimento, anormalidade morfológica, entidade de observação, estrutura corpórea, evento, valor qualificador, substância, objeto físico, força física, transtorno, contexto social, artefato de armazenamento, achado clínico e conceitos de ligação.

Na figura 3 é demonstrada com a hierarquia adotada e as instâncias da classe "Procedimento".

FIGURA. 3
AS CLASSES E INSTÂNCIAS DA CLASSE PROCEDIMENTO NO PROTÉGÉ.

Atualmente nosso trabalho está em criar os documentos compatíveis ao padrão HL7 CDA de tal forma a criar uma forma possível de troca de informações entre sistemas, utilizando arquivos eletrônicos de documentos médicos, utilizando este modelo internacional.

CONCLUSÃO

O estudo apresentado demonstra uma metodologia de extração de conhecimento de laudos de radiologia torácica, escrito em linguagem natural, que possam ser reutilizados na troca de informações entre sistemas, orientando engenheiros biomédicos e profissionais que necessitem desse conhecimento.

O estudo de processamento estatístico de linguagem natural não ofereceu resultados diferentes daqueles identificados em estudos utilizando a língua inglesa, visto que o estudo da língua decorre de uma análise simbólica da expressão escrita.

Na análise semântica a maioria das palavras e *n-gramas*, não foi encontrada na classificação do Snomed-CT. Foram 1.067 palavras e trechos de palavras identificadas no processamento de linguagem natural e 361 palavras havia conceitos correspondentes ao Snomed-CT.

Analisando os *n-gramas*, que são frases ou segmento de frases o resultado não foi diferente. Dos 836 *n-gramas* identificados, 160 tem uma correspondência aos conceitos do Snomed-CT.

Esse desempenho decorre também da base SNOMED-CT utilizada, que somente padroniza os termos clínicos, não classificando vários termos de abrangência mais ampla no conhecimento médico.

Como poucas palavras correspondem a grande percentual de ocorrência em um corpus, as poucas palavras encontradas apresentam uma significativa parte dos conceitos de radiologia dos dois mil laudos analisados, mas não abrange a sua totalidade.

Nota-se também que as frases que os digitadores e médicos selecionam, no sistema de edição de laudos de raios-X de tórax, possivelmente não tiveram uma análise prévia de correspondência à terminologia clínica adotada pela comunidade internacional, acarretando o conhecido problema de terminologias médicas com termos diversos que podem representar o mesmo conceito, dificultando a busca de informações de casos similares e o estudo estatístico de bases de conhecimento baseados em texto.

O próximo passo deste trabalho será a criação dos documentos ora representados em ontologias, utilizando o OWL como linguagem, em documentos compatíveis com o HL7 CDA.

AGRADECIMENTOS

Agradeço ao pessoal do Incor pelo fornecimento dos laudos de radiologia torácica e pelas informações sobre o processo de preenchimento dos laudos.

REFERÊNCIAS

[1] Massad, Eduardo e Rocha, Armando Freitas. O Prontuário Eletrônico do Paciente na Assistência, Informação e Conhecimento Médico. *Livro*. 2003.

[2] Gordon, Raymond G e Grimes, Barbara F. Ethnologue: languages of the world. *Livro*. 2005.

[3] Gough, Kevin J. Syntax analysis and software tools. *Livro.* 1988.

[4] Manning, Christopher D e Schütze, Hinrich. Foundations of Statistical Natural Language Processing. *Livro.* 1999

[5] Almeida, Maurício B e Bax, Marcello P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Revista Ciência da Informação*, 2003.

[6] Chaui, Marilena. Convite à Filosofia. *Livro*. 2003

[7] Casare, Sara J. Uma ontologia funcional de reputação para agentes. *Dissertação*. 2005.

[8] Borst, Pim, Akkermans, Hans e Top, Jan. Engineering ontologies. *International Journal of Human Computer*. 1997.

[9] Huang, Yang, et al. Improved Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with UMLS Specialist Lexicon. *American Medical Informatics*. 2005.

[10] Meystre, Stephan e Haung, Peter J. Medical Problem and Document Model for Natural Language Understanding. *American Medical Informatics.* 2003.

[11] Hunscher, Dale, et al. Representing Natural-Language Case Report Form Terminology Using Health Level 7 Common Document Architecture , LOINC, and SNOMED-CT: Lessons Learned. *American Medical Informatics*. 2006.