Edson José Pacheco[1,2]
Percy Nohama[1]
Stefan Schulz[3]

Model for encoding clinical narrative into a domain ontology

Codificação de narrativas clínicas para uma ontologia de domínio

ABSTRACT | Introduction: *The increasing expansion of medical information available in health records and biomedical literature reflects the need to generate processes that facilitate the retrieval of documents and information. In this context exists the encoding process that uses controlled vocabularies to describe this information and aims at efficient use of search tools.* **Objective:** *In this paper, we present a set of rules for coding and structuring of clinical texts to SNOMED CT, basis for the automatic mapping process.* **Methods:** *for the experiments, we select a set of clinical documentation from the cardiology department of a Brazilian hospital.* **Results:** *A hundred and twenty clinical discharges were coded, 12 was mappead by more than one expert, resulting in 89% of agreement (Kappa).* **Conclusions:** *It was concluded, therefore, the effectiveness of the developed guideline, which will serve as the basis for the automatic mapping process.*

Keywords | *Knowledge management for health research; Systematized Nomenclature of Medicine; Natural language processing.*

RESUMO | Introdução: A crescente expansão de informações médicas disponíveis em prontuários e literatura biomédica reflete a necessidade da geração de processos que facilitem a busca de documentos e informações. Nesse contexto, insere-se o processo de codificação que usa termos controlados para descrever essas informações e objetiva a eficácia na utilização de ferramentas de busca. **Objetivo:** No presente artigo, propõe-se um conjunto de regras (*guideline*) para codificação e estruturação de textos clínicos para a *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED CT), oportunizando a constituição de algoritmos computacionais. **Métodos:** Para os experimentos, foram utilizados um conjunto de prontuários oriundos do Departamento de Cardiologia de um hospital brasileiro. **Resultados:** Foram codificados 120 sumários de alta, dos quais 12 foram mapeados por mais de um profissional, o teste Kappa obtido foi de 89%, indicando um elevado grau de concordância entre os profissionais. **Conclusão:** Conclui-se, portanto, pela efetividade do *guideline* desenvolvido, o qual servirá de base para o processo de mapeamento automático.

Palavras-chave | Gestão do conhecimento para a pesquisa em saúde; Systematized Nomenclature of Medicine; Processamento de linguagem natural.

[1]Pontifícia Universidade Católica do Paraná. Curitiba/PR, Brasil.
[2]Universidade Federal do Paraná. Curitiba/PR, Brasil.
[3]Medizinische Universität Graz. Graz/Steiermark, Áustria.

INTRODUÇÃO |

Os estudos antropológicos, tanto na linha evolucionária quanto na do desenvolvimento filogenético, mostram que os homens têm uma tendência "espontânea" a "descobrir" o que é o mundo que os circunda, a conhecer, a buscar compreender o que é este mundo[1,2,3].

O conhecimento, nesse contexto, configura-se como o esforço do "espírito" humano para entender a realidade circundante. Essa compreensão ocorre mediante a interpretação dos sinais captados pelos sentidos. O significado, a partir da interpretação dos sinais, é resultante dos processos cognitivos que, por sua vez, acontecem mediante a associação entre símbolos representando objetos do mundo e a realidade observada. O elemento central dessa subjetividade instauradora de relações entre os diversos aspectos da "realidade" é exatamente a capacidade de "duplicar" elementos da experiência humana pelo processo de simbolização (representação mental), permitindo que a realidade circundante seja tratada no plano simbólico[3,4].

Como representação primeira, a linguagem é o meio mais utilizado para a representação e disseminação de conhecimento, elemento essencial no processo de simbolização. No domínio da disseminação, o primeiro grande paradigma foi o da dialética, substituído por novas técnicas e metodologias apoiadas na antiga, visto que a escrita nasce com a consciência humana e nova forma, no aspecto da facilidade na replicação de registrar o circundante, quando do advento da impressão e tipografia, possibilitado pela invenção de Johannes Gutenberg[5].

Com a definição das bases, técnicas e conceituais, de comunicação e intercâmbio de informações e conhecimento, os últimos séculos foram caracterizados pelo acúmulo desses elementos, permitindo a construção de bibliotecas, de métodos de indexação para a localização da informação desejada. Em síntese, com o crescimento da disponibilidade, surge a necessidade da guarda e recuperação da informação[6,7].

Com a adoção das tecnologias de informação, tais mecanismos, os de guarda, representação e recuperação, sofreram profunda alteração nos últimos anos, propiciando a produção e manipulação de grandes volumes de dados, informações e conhecimento, virtualmente disponíveis a todos pelo uso das tecnologias de acesso[8].

Na área médica, complexidade adicional é identificada dado que as bases de documentos médicos e de dados clínicos são extensas e dinâmicas, evidenciado a necessidade de manuseio de terminologias clínicas.

A utilização de processos de codificação pode incrementar a eficiência da comunicação eletrônica e o processamento das informações clínicas, melhorando a qualidade dos dados armazenados. Além disso, a criação de regras para codificação manual tem grande relevância, uma vez que é importante que o mapeamento realizado por diferentes profissionais (atendimento multiaxial e multiprofissional) tenha o mesmo direcionamento.

Dentre as terminologias de domínio que abrangem a área da saúde, a SNOMED CT (*Systematized Nomenclature of Medicine - Clinical Terms*) apresenta, atualmente, maior abrangência[9]. A SNOMED CT é um vocabulário médico padronizado que define o significado dos termos médicos, visando a melhorar a interoperabilidade entre sistemas e inclui sinais, sintomas, diagnósticos e procedimentos[10]. Cabe destacar a adoção da terminologia pelo Ministério da Saúde para "[...] codificação de termos clínicos e mapeamento das terminologias nacionais e internacionais em uso no país, visando suportar a interoperabilidade semântica entre os sistemas" [11].

O objetivo da pesquisa descrita neste artigo é propor uma metodologia para o mapeamento de texto livre, especificamente de narrativas clínicas, para a SNOMED CT, por meio de regras de codificação.

MÉTODOS |

Snomed CT

A revisão constante da SNOMED CT, com a inclusão de relações ontológicas, permite classificá-la como uma ontologia em processo de revisão e concepção ou, ainda, sob uma linha conceitual menos estrita, como uma ontologia propriamente dita.

A SNOMED CT é composta por um conjunto de conceitos, termos e relações que objetivam a precisa representação de informação clínica, no domínio da saúde. A cobertura da terminologia é dividida em hierarquias, objetivando facilitar a identificação do conhecimento representado.

Os elementos básicos que a compõem são[12]:

a) conceitos: unidades básicas da SNOMED CT que representam uma "unidade de significado". Um conceito é definido por um código numérico único, nome único (*Fully Specified Name*), um conjunto de termos (*descriptions*), um "*Preferred Term*" e sinônimos;

b) descrições: termos ou nomes atribuídos a um conceito;

c)hierarquias: são 19 as principais (ex.: *Disorder, Procedure, Substance, Clinical findings,* etc.);

d) relações: associações entre conceitos, frequentemente usadas para formalmente definir um conceito. Exemplo:

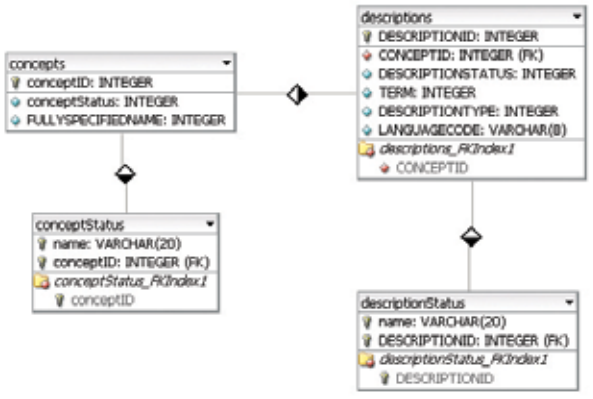
Fully Specified Name: Fracture of femur (disorder)
Is a: Fracture of lower limb (disorder)
Is a: Injury of thigh (disorder)

Group 1
Associated morphology: Fracture (morphologic abnormality)
Finding site: Bone structure of femur (body structure)

O exemplo expressa: uma fratura de fêmur é uma fratura do membro inferior (*Fracture of lower limb)* e um ferimento na coxa (*Injury of thigh)*. Além disso, *Fracture of femur (disorder)* tem relação com dois conceitos: fratura (*Fracture (morphologic abnormality))* e estrutura óssea do fêmur (*Bone structure of femur).*

A SNOMED CT é distribuída sobre a forma de arquivo-texto, *flat*, com quatro artefatos principais, relacionados segundo o modelo apresentado na Figura 1. Os arquivos são distribuídos separadamente de acordo com o idioma. Cada distribuição (conjunto de arquivos que compõem a SNOMED CT) é disponibilizada de acordo com um modelo de licenciamento específico. As distribuições são qualificadas com o mês e ano.

Figura 1 - Modelo relacional da distribuição da SNOMED CT



As estruturas "conceptStatus" e "descriptionStatus" são qualificadoras dos registros relacionados, seja de conceitos, seja de descrições.

O conjunto de valores admitidos para "conceptStatus", na distribuição considerada, é: corrente, retirado, duplicado, desatualizado, ambíguo, errado e limitado.

O conjunto de valores admitidos para "descriptionstatus", para a mesma distribuição, é: corrente, não corrente, duplicado, desatualizado, errado, limitado, inapropriado, movido para outra estrutura e movimentação pendente.

Os domínios para "*descriptionType"*, na tabela "*description*" são:

a) *Preferred Terms*: denota uma descrição padrão para o conceito relacionado*;*

b)*FullySpecified Name:* descrição composta por um dos *Prefered Terms* e a hierarquia de nível mais alto*;*

c)*Explanation or Definition*: definição*;* e

d) *External reference*: definição importada para a SNOMED CT.

O conjunto domínio para o campo "*LANGUAGECOD"*, da tabela "*description"*, para os idiomas espanhol e inglês, é:

a) en: descrições comuns para o idioma inglês;

b) en-GB: descrições específicas para o inglês britânico;

c) en-US: descrições específicas para o inglês americano, e

d) sp: descrições para o idioma espanhol.

Base de estudo

Na pesquisa em questão, envolvendo a codificação de textos clínicos para a SNOMED CT, optou-se pelo uso de sumários de alta. A escolha desse tipo de documento deu-se pela sua abrangência, multidisciplinaridade e agregação das informações médicas do paciente. Um sumário de alta contém, ao menos: sinais e sintomas do paciente, antecedentes pessoais e familiares, exame físico, laudos, medicações usadas e planos para o seguimento do caso[13].

Os sumários utilizados foram obtidos no Hospital de Clínicas de Porto Alegre (HCPA), tendo sido selecionada a área de cardiologia para a concretização dos estudos, contemplando o período de junho de 2002 a maio de 2007.

Construção de *corpus* anotado para a avaliação linguística de documentos médicos

O Processamento de Linguagem Natural (PLN) é uma subárea da IA e daLinguística, tendo como objeto a extração de informações computáveis a partir de textos. Os algoritmos de PLN demandam intenso processamento computacional[14], principalmente pelas características da linguagem natural, entre essas:

a) rica e elaborada e ao mesmo tempo vaga e ambígua;

b) os significados dos termos são, ao mesmo tempo, independentes e associados a outros termos, e

c) há inúmeras formas de se dizer a mesma coisa.

Tradicionalmente, a primeira etapa do processamento da linguagem natural é a delimitação das sentenças e a identificação dos *tokens*. Posteriormente, as estruturas associadas a cada item, como: gênero e número para substantivos ou pessoa e número, modo e tempo para os verbos[15]. Isso, porém, requer que as palavras possam ser identificadas mediante um léxico que também implementa o conhecimento da formação das palavras.

O analisador léxico-morfológico e o etiquetador gramatical são "interconectados". O segundo é o efetivamente responsável pela etiquetagem, para cada item lexical, da categoria a que esse item pertence. A etiquetagem é o "[...] processo de demarcação de um marcador de classe gramatical (ou outro marcador ou 'etiqueta' de interesse) a cada palavra, num *corpus*" [15].

Os etiquetadores podem ser constituídos baseados em regras ou implementados pelo modelo estocástico. No primeiro, regras são definidas de forma a permitir a identificação da categoria de um item lexical. A desvantagem desse mecanismo é que novas regras são manualmente adicionadas quando da identificação de novas situações. No caso do modelo estocástico, a estratégia de ação é iniciada pelo treinamento baseado em um *corpus* previamente marcado. Pelo cálculo de probabilidade que um determinado item terá para cada etiqueta disponível.

O analisador sintático contempla a etapa seguinte no PLN, objetivando o reconhecimento de uma sequência de palavras como uma sequência válida, ou não, da língua considerada. A análise sintática completa é dificultada pelo surgimento de um alto número de ambiguidades semânticas, assim como pela complexidade dos algoritmos. Na prática, o "*parser*" é um processo que analisa uma sequência de entrada objetivando a identificação da estrutura gramatical, segundo um determinado formalismo. Procura-se, nesse cenário, chegar pelo menos a uma delimitação adequada de frases ou até entidades menores (*chunks*).

De forma complementar ao analisador sintático, Morton (2006)[15] destaca o detector de frases nominais, que são aquelas que encerram em si significado completo, estático e independente.

Considerando a natureza nominativa dos conceitos estruturados na SNOMED CT, buscou-se a construção de uma base anotada, *corpora*, no domínio de aplicação (área médica), para o treinamento do *framework* de PLN OpenNLP[16]. A OpenNLP é uma solução *open source* e inclui ferramentas que englobam todas as etapas relacionadas com a PLN necessárias para o desenvolvimento do presente trabalho.

Para o processo de etiquetagem morfológica, foi utilizada a metodologia de "*Active Learning*"[17], que se baseia no uso de um comitê de etiquetadores — cada qual utilizando métodos de análise específicos ou comuns — para avaliar as sentenças com maior grau de discordância e que, portanto, necessitam de avaliação manual.

Segundo Tomanek, Wermter e Hahn[17], a estratégia é capaz de reduzir em até 50% o número de palavras a serem etiquetadas por humanos se comparada com uma técnica de seleção aleatória, alcançando os mesmos índices de exatidão.

Essa metodologia implica duas decisões centrais, anteriores ao processo de desenvolvimento. A primeira consiste na escolha do comitê de etiquetadores. Nesse sentido, optou-se pelas ferramentas *Lácio-Web — MXPOST*[1], *TreeTagger*[2] e *Brill Tagger*[3] —, pelo etiquetador *QTag*[4] e pela *OpenNLP*. O comitê de etiquetadores foi selecionado considerando a exatidão mínima superior a 90%, quando aplicados em textos jornalísticos.

A segunda decisão importante refere-se ao conjunto de etiquetas que terá prioridade sobre os demais, para o qual todas as saídas devem ser mapeadas. O conjunto selecionado para o domínio de aplicação é apresentado no Quadro 1, com prioridade decrescente.

Para a elaboração do *corpus* foram escolhidos aleatoriamente 4.564 sumários de alta dentre os cedidos para estudo, perfazendo um total de 622.255 *tokens*, nominado de *corpus* de TREINAMENTO.

O estado inicial foi construído com base no *corpus* de TREINAMENTO e treinado com as ferramentas disponíveis

na OpenNLP[16], com base no modelo produzido a partir do Mac-Morpho *Corpus*[5]. O etiquetador morfológico da OpenNLP é baseado no algoritmo de Viterbi para modelos de Markov de segunda ordem. O *corpus* de TREINAMENTO devidamente etiquetado é nominado de T.

Quadro 1 - Etiquetas utilizadas

Classe Gramatical	Etiqueta
Pronome	PRN
Nome próprio	NPROP
Substantivo ou adjetivo	NADJ
Numeral	NUM
Advérbio	ADV
Artigo	ART
Conjunção	CJ
Preposição	PREP
Palavra denotativa	PDEN
Particípio	PCP
Interjeição	IN
Verbo	V
Símbolo de moeda corrente	CUR

A partir da definição do estado inicial, seguem-se iterações de treinamento com cada uma das ferramentas que compõem o comitê de etiquetadores, com base em T2. Para cada sentença de TREINAMENTO, é realizada a etiquetagem individual pelos membros do comitê de etiquetadores, permitindo a seleção de sentenças, com maior representatividade no aumento da exatidão dos etiquetadores, objetivando a correção manual. O fluxo de trabalho, que é concluído num intervalo de 24 horas, é ilustrado na Figura 2.

Figura 2 - Fluxo de trabalho com o Active Learning

A seleção de sentenças é realizada por meio do cálculo do índice de discordância de sentença $D_{sent}(s)$, função do índice de discordância de *token* $D_{tok}(t)$, para cada sentença sem marcação manual em uma determinada iteração. Tais índices expressam, em uma escala variável de 0 a 1, o grau de inconsistência de resultados obtidos pelo comitê de etiquetadores, ora em nível de *token*, ora em nível de sentença. As equações "Equação 1" e "Equação 2" (Tomanek, Wermter e Hahn, 2007)[17], expressam matematicamente esse propósito. Nelas, $\frac{V(l_i,t)}{k}$ é a razão de k etiquetadores que atribuíram a etiqueta l_i para um token t e $|s|$ é o tamanho da sentença sob análise.

$$D_{tok}(t) = -\frac{1}{\log k}\sum_{l_i}\frac{V(l_i,t)}{k}\log\frac{V(l_i,t)}{k}$$ Equação1

$$D_{sent}(s) = \sum_{j=1}^{|s|}\frac{D_{tok}(t_j)}{|s|}$$ Equação 2

A cada iteração de treinamento, o modelo criado para a *OpenNLP* é avaliado quanto à exatidão sobre o conjunto VALIDAÇÃO. O valor de incerteza encontrado em cada iteração e a média dos índices de discordância $D_{sent}(s)$ $D_{sent}(s)$ das sentenças sem marcação manual são armazenados para posterior análise gráfica de suas variações em função do número de *tokens* corrigidos a cada passo.

A etapa de correção manual é realizada por especialistas de domínio, com apoio da ferramenta desenvolvida especificamente para esse propósito (Figura 3). Na ferramenta em questão, é apresentado ao especialista um conjunto de sentenças para correção, em que a etiqueta predominante (mais frequente dentre as sugeridas pelo comitê de etiquetadores) para cada *token* é destacada e sugerida como correta, permitindo a correção manual no caso de indicação errônea. A ferramenta ainda exibe de forma gráfica os valores de exatidão e a média dos índices de discordância obtidos nas iterações antecedentes, além de destacar o último valor obtido.

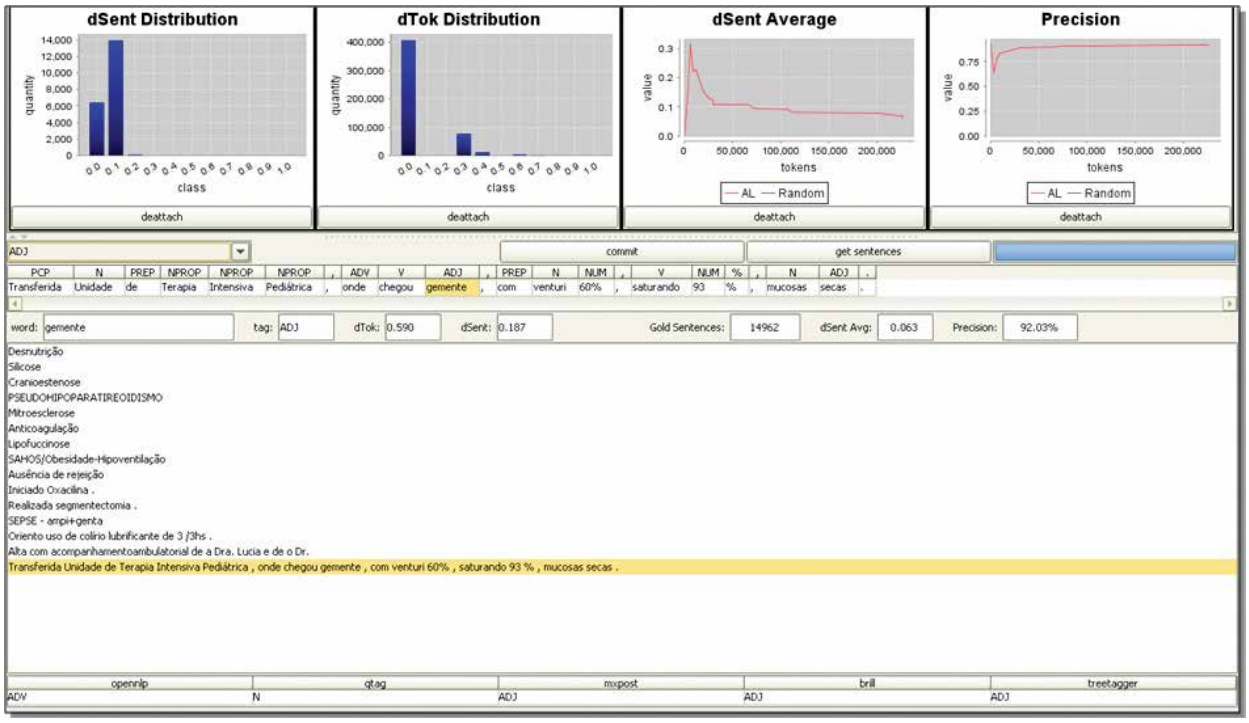
O ciclo foi interrompido quando do alcance de um índice estacionário de exatidão[6], após 160 iterações completas. A precisão alcançada é de 93,67%, compatível com iniciativas similares para outros idiomas[25].

Concluída a etiquetagem morfológica, o corpora gerado foi utilizado para o treinamento do detector de frases nominais da OpenNLP.

Avaliação e levantamento de regras de mapeamento

Para a identificação de padrões de mapeamento, foram utilizados 1.200 sumários de alta, selecionados de maneira

Figura 3 - Interface gráfica da ferramenta de correção manual de etiquetas



aleatória, garantindo-se que nenhum dos selecionados tenham sido utilizados na etapa de etiquetagem morfológica.

Visando a uma melhor padronização do estudo, os sumários foram convertidos em uma estrutura tabular. A divisão do texto realizou-se de tal maneira que cada linha da tabela correspondia a um *chunk*, isto é, um termo médico ou um "não-termo", como um verbo, advérbio, conjunção, identificados por meio da OpenNLP.

Foram adicionadas aos *chunks* colunas paralelas para serem usadas como interpretação do prontuário sem interferir na codificação. As colunas adjacentes foram denominadas de *Context, Polarity* e *Observation*.

Na coluna *Context*, foram aplicadas dez interpretações possíveis:

1 - Stopped (STP): refere-se a algum processo que terminou no período do internamento (ex.: "a medicação foi suspensa") ou final de sinais e sintomas (ex.: "remissão");

2 * *History* (HIS): refere-se a condições que ocorreram antes do internamento;

3 - *Family history* (FAM): refere-se a condições presentes em familiares do paciente;

4 - *Uncertainty* (UNC): refere-se a um enunciado incerto. Ex.: "disfunção respiratória, provavelmente causada por aspiração";

5 - *Plan* (PLA): refere-se ao plano de alguma intervenção. Ex.: "a implantação de um stent foi planejada";

6 - *Imperative /Optative* (IMP): expressão de desejo ou ordem para fazer algo. Ex.: "o *stent* deve ser implantado";

7 - *Hypothetic* (HYP): expressão do pensamento sobre alguma coisa. Ex.: condicional ("o stent deveria ter sido implantado");

8 - *ide effects* (SEF): expressão de efeito não desejado de alguma droga ou terapia;

9 - *Necessity* (NEC): um procedimento é considerado necessário;

10 - *Risk* (RSK): risco de sintoma ou doença.

A coluna *Polarity* é atribuída à abreviatura NEG (negação) e foi adicionada na existência de uma partícula de negação ou um verbo com significado negativo.

Na última coluna, *Observation*, quando necessário, incluiu-se algum comentário do *chunk* correspondente. Essa coluna foi usada principalmente para decifrar os acrônimos existentes nos prontuários médicos.

As características elaboradas para facilitar a codificação e análise dos termos relevantes contidos nos sumários de alta são demonstradas no Quadro 2.

Como exemplo de um fragmento de prontuário, ilustra-se: *"Paciente com história de HAS e DM tipo 2 interna por angina aos grandes esforços. Submetido a cat que evidenciou lesão suboclusiva na coronária DA. Realizado angioplastia. Recomendações na alta: retorno ao seu médico e em caso de intercorrência vir à emergência"*.

O processo de codificação manual de termos clínicos contidos nos sumários de alta foi realizado por dois profissionais da área da saúde.

A concretização do trabalho abrangeu uma série de etapas, as quais impulsionaram a composição de um *guideline,* que foi implantado para dar suporte à criação de um padrão ouro de codificação objetivando a avaliação de um processo computadorizado de mapeamento, usando tecnologias de processamento de linguagem natural.

Finalmente, foram criadas regras para diminuir a ocorrência de divergências na codificação:

a) Não use conceitos SNOMED CT que expressem negações, por outro lado codifique o conceito e adicione a coluna Polarity: NEG.

b) Não use conceitos SNOMED CT contexto-específico (ignore a hierarquia "*Context dependent categories*").

c) Somente use a hierarquia *"Morphologies*", se não houver código correspondente nas hierarquias preferidas: "*Clinical Findings*" ou "*Procedure*".

d) Para codificar resultados de laboratório, utilize a categoria "O*bservable concepts*";.

e) Para codificar um medicamento, utilize o código SNOMED CT referente à "*Substance*" ao invés de "*Product*". Ex.: "*digoxin (substance)*" preferencialmente à "*digoxin (product)*".

f) Se marcas comerciais forem usadas nos texto clínicos, codifique o nome genérico da substância.

Quadro 2 – Exemplo da estrutura tabular utilizada no estudo

Chunk	Context	Structure	Polarity	Code	Observation
Paciente					
com história de					
HAS e	HIS	LIS		38341003	Hipertensão arterial sistêmica
DM tipo 2	HIS	LIS		44054006	Diabetes mellitus
interna por					
angina aos grandes esforços.				300995000	
Submetido a					
Cat				41976001	cateterismo
que evidenciou					
lesão suboclusiva				2929001	
na coronária DA.				59438005	Descendente Anterior
Realizado					
angioplastia.				41339005	
Recomendações na alta:					
retorno ao	PLA	LIS			
seu médico e	PLA	LIS			
em caso de intercorrência	UNC	LIS			
vir à	UNC	LIS			
emergência.	UNC	LIS		185265000	

g) Existem conceitos SNOMED CT diferentes para cada especialidade médica. Ex.: *Cardiology service (procedure), Seen by cardiologist (finding), Cardiologist (occupation)*. Nesses casos, opte por codificar o termo que reflete a ocupação (*Cardiologist (occupation)*) para todas as especialidades.

h) Para conceitos anatômicos, a SNOMED CT apresenta sempre dois conceitos: "*Entire X*" e " *X structure* ". Nesses casos, codifique o último, termo mais geral. Ex.: *"Right coronary artery structure"* ao invés de *"Entire right coronary artery"*.

i) Utilize conceitos SNOMED CT da categoria "*Qualifier value*" apenas quando esta referir-se a características importantes de outros conceitos. Ex.: "*laterality, episodicity, severity*".

j) Se um termo estiver presente na hierarquia SNOMED CT "*Anatomical concepts*" e o mesmo conceito estiver nas hierarquias "*Clinical Findings*" ou " *Procedure*", codifique o termo obtido na " *Clinical Findings*".

k) Evite conceitos que refinam outro conceito por informações diagnósticas. Ex.: "*Invasive carcinoma of uterine cervix diagnosed by microscopy only*".

l) Evite usar conceitos "híbridos" (conceitos que misturam a situação do paciente com a ação de observação pelo médico). Ex.: Para codificar que um paciente está com febre, prefira utilizar o código "*Fever (finding)*" ao invés de " *On examination – fever (finding)*".

RESULTADOS/DISCUSSÃO |

A efetividade do *guideline* produzido foi conferida pela avaliação da concordância do trabalho de codificação realizada pelos profissionais. Para isso, 300 sumários de alta foram aleatoriamente selecionados e submetidos ao processo de codificação pelos dois especialistas envolvidos nessa atividade. Para a obtenção do resultado, utilizou-se o teste Kappa (k), que é o procedimento estatístico utilizado para avaliar a confiabilidade de variáveis categóricas e nominais[20]. Kappa é interpretado como a proporção de concordância entre duas ou mais medidas de n observações, após a exclusão das concordâncias ao acaso. Com base na análise do mapeamento gerado, o Kappa obtido foi de 89%, indicando um elevado grau de concordância entre os profissionais.

Por meio desse trabalho, foram obtidos os casos em que houve maior discordância na codificação e assim foi possível observar que alguns conceitos SNOMED CT têm similaridade. Um exemplo muitas vezes presente nos textos clínicos é o aparecimento da palavra "tabagista", cuja codificação foi diferente entre os profisssionais. Os códigos obtidos foram correspondentes a: *Tobacco user* (*finding*) e *Smoker (finding)*. Na SNOMED CT, *Tobacco user* (*finding)* é um conceito que engloba *Smoker* (*finding)* e outros conceitos como: *Chews tobacco (finding), Snuff user (finding)* e *Use of moist powdered tobacco (finding)*. Portanto, a escolha do código para a palavra "tabagista" foi dificultada pelas nuanças dos conceitos presentes na SNOMED CT e pela semelhança entre eles.

O conceito para a palavra "eletrocardiograma" tem duas interpretações possíveis: *Electrocardiogram finding (finding) e Electrocardiographic procedure (procedure)*. Assim sendo, para que não houvesse divergência na codificação, seria necessária a interpretação do contexto clínico. Na SNOMED CT, *Electrocardiogram finding (finding)* apresenta como sinônimo: *ECG observations*, então, nos casos em que estivesse presente alguma observação do eletrocardiograma, este teria que ser interpretado como *Electrocardiogram finding (finding)*, como no exemplo: "presença de ondas Q no eletrocardiograma". Já *Electrocardiographic procedure (procedure)* ficaria reservado para ocasiões em que não contivesse observação do eletrocardiograma, exemplo: "realização de eletrocardiograma".

O termo "insuficiência renal crônica (IRC)" obteve dois códigos que correspondem aos conceitos: *Chronic renal impairment (disorder)* e *Chronic renal failure syndrome (disorder)*. Na SNOMED CT, *Chronic renal impairment (disorder)* é um conceito que engloba *Chronic renal failure syndrome (disorder)* e outros conceitos como *Chronic kidney disease stage 1, Chronic kidney disease stage 2, Chronic kidney disease stage 3, Chronic kidney disease stage 4, Chronic kidney disease stage 5. Chronic renal failure syndrome (disorder)* apresenta, como sinônimo na SNOMED CT, o conceito *CRF - Chronic renal failure*, ou seja, o conceito equivalente para o termo proposto (insuficiência renal crônica (IRC)). Entretanto, na bibliografia médica, a insuficiência renal crônica (IRC) apresenta cinco estágios e, na SNOMED CT, os conceitos desses estágios estão contidos no conceito de *Chronic renal impairment (disorder)* e não no conceito *Chronic renal failure syndrome (disorder)*. Portanto, houve dificuldade em enquadrar o termo "insuficiência renal crônica (IRC)" em um dos conceitos apresentados pela SNOMED CT.

Na maioria das ocorrências de termos clínicos, foi evidenciada a concordância entre os profissionais envolvidos, ilustrando a viabilidade da metodologia proposta. Algumas situações específicas, associadas à estrutura da SNOMED CT, foram

identificadas no desenvolvimento do trabalho, dentre as quais é relevante citar:

a) O conceito para "disfunção de ventrículo esquerdo" teve os códigos correspondentes a: *Left ventricular cardiac dysfunction (disorder)* e *Disorder of left cardiac ventricle (disorder)*. Embora *Left ventricular cardiac dysfunction (disorder*) seja uma tradução fiel para o conceito proposto, *Disorder of left cardiac ventricle (disorder)* também pode ser interpretado como uma "disfunção de ventrículo esquerdo".

b) Na codificação para o termo "saturação de oxigênio", foram obtidos os seguintes conceitos na SNOMED CT: *Oximetry (procedure)* e *Oxygen saturation measurement, arterial (procedure)*. Ambos são conceitos sinônimos, o que ocasionou duas codificações possíveis para um mesmo termo.

c)Para o termo "parada respiratória", os conceitos SNOMED CT foram: *Stops breathing (finding)* e *Respiratory arrest (disorder)*. Ambos têm o mesmo significado, entretando um relaciona-se com a hierarquia *Clinical findings* e outro com a *Disorder*. No *guideline* produzido, não foi proposta a preferência para codificação entre essas duas hierarquias em caso de haver um mesmo conceito presente em ambas, o que gerou a divergência na codificação.

d) A palavra "hematúria" obteve os seguintes conceitos SNOMED CT: *Hematuria syndrome (disorder)* e *Blood in urine (finding)*. *Hematuria syndrome (disorder)* apresenta como sinônimo na SNOMED CT: *Blood in urine – hematuria*. Portanto, há dois conceitos SNOMED CT com o mesmo significado, o que dificultou a concordância na codificação.

Por intemédio deste trabalho, também foi possível desenvolver o conhecimento das hierarquias que mais obtiveram concordância no processo de codificação:

a) medicamentos (Hierarquia *Substance*): Aspirina, Sinvastatina, Metoprolol, Captopril, Furosemida, Hidroclorotiazida etc.;

b) doenças (Hierarquia *Disorder*): hipertensão arterial sistêmica, diabetes *mellitus*, cardiopatia isquêmica, insuficiência cardíaca congestiva, infarto;

c) procedimentos (Hierarquia *Procedure*): angioplastia transluminal percutânea, raios x de tórax, ecocardiograma, cintilografia miocárdica;

d)estruturas anatômicas (Hierarquia *Body Structure*): ventrículo direito, septo cardíaco, parede posterior do coração, artéria coronária direita, artéria circunflexa, artéria descendente anterior.

CONCLUSÃO|

Representar o conhecimento armazenado em narrativas clínicas é objeto de estudo recente, especialmente pelo intenso investimento no século XX, no sentido de estabelecer as bases conceituais e tecnológicas para o registro de eventos clínicos[21,22], produzindo grandes bases de dados e conhecimento que, agora, exigem novas técnicas para a sua manipulação.

O mapeamento de artefatos para estruturas formais tem recebido atenção recente da comunicação científica, com a publicação de trabalhos relevantes na área[23,24,25]. Como elemento diferenciador, o presente trabalho trata do processo de recuperação e mapeamento contemplando todo o ciclo: do registro do conhecimento (texto livre), passando pelo mapeamento e representação da estrutura conceitual implícita.

Apesar das dificuldades encontradas na codificação de alguns termos pela existência de conceitos SNOMED CT semelhantes, o *guideline* proposto abordou tópicos importantes para a obtenção de um resultado satisfatório de codificação manual, o que representa um passo importante em direção à informatização eficiente de documentos médicos.

O Kappa de 89% é indicativo relevante da adequação metodológica do trabalho realizado.

REFERÊNCIAS|

1 - Konder L. A revanche da dialética. São Paulo: UNESP; 2002.

2 - Jiang M, Cui P, Liu R, Yang Q, Wang F, Zhu W, et al. Social contextual recommendation. Proceedings of the 21st ACM international conference on Information and knowledge management; 2012 Oct/Nov 29-02; Maui, USA. New York: CIKM; 2012.

3 - Severino, A. Síntese do conhecimento. 23 ed. São Paulo: Cortez Editora; 2002.

4 - Viale R. Cultural and cognitive dimensions of innovation. Mind Soc. 2013. 12:1-3.

5 - Liu Q, Tan CC, Wu J, Wang G. Efficient information retrieval for ranked queries in cost-effective cloud environments. Proceedings of the 31st IEEE International Conference on Computer Communications; 2012 Mar 25-30; Orlando, USA. New York: IEEE Communications Society; 2012.

6 - Marques A. Para entender as linguagens documentárias. 2 ed. São Paulo: Polis; 2002.

7 - Petit M, Lallee S, Boucher JD, Pointeau G, Cheminade P, Ognibene D *et al*. The coordinating role of language in real-time multimodal learning of cooperative tasks. Autonomous Mental Development, IEEE Transactions. 2013; 5(1): 3-17.

8 - Tran DH, Nguyen HP. API specification-based function search engine using natural language query. Proceedings of the International Conference on Computing, Management and Telecommunications (ComMaTel); 2013 Jan 21-24; Ho Chi Minh City, Vietnam. Washington: IEEE; 2013.

9 - International Health Terminology Standards Development Organization [Internet]. Copenhagen: IHTSDO [citado 2013 maio 9]. Disponível em: http://www.ihtsdo.org/snomed-ct.

10 - Alexandrini F, Vermöhlen J, Cattoni AA. Prontuários eletrônicos de pacientes em padrão DICOM-SR/HL7. Revista Caminhos. 2006; 7(1):175-95.

11 - Brasil. Ministério da Saúde. Portaria nº 2.073, de 31 de agosto de 2011. Diário Oficial da República Federativa do Brasil, Brasília, 01 set 2011, col 1, p.63.

12 - Rogers J, Bodenreider O. SNOMED CT: browsing the browsers. In: Cornet R, Spackman KA, editors. *Representing and sharing knowledge using SNOMED*. Proceedings of the 3rd international conference on Knowledge Representation in Medicine; 2008 May 31-June 02, Phoenix, USA. Copenhagen: IHTSDO; 2008. p. 30-36.

13 - Kluck MM, Guimarães JR. Sumário eletrônico de alta: garantindo a continuidade da assistência ao paciente através da informação. Informática Pública. 1999; 1(2):123-37.

14 - Jones D, Somers H. New methods in language processing. London: University College Press; 2011.

15 - Morton T. Using semantic relations to improve information retrieval [dissertation]. Philadelphia: University of Pennsylvania; 2005.

16 - Apache OpenNLP [Internet]. Los Angeles: The Apache Software Foundation; c2010 [citado 2013 maio 09]. Disponível em: http://opennlp.sourceforge.net/.

17 - Tomanek K, Wermter J, Hahn U. An Approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational; 2007 June; Prague. Stroudsburg: Association for Computational Linguistics; 2007.

18 - Kuncheva LI. A Bound on kappa-error diagrams for analysis of classifier ensembles. Knowledge and Data Engineering. 2013; 25(3):494-501.

19 - Aluisio SM, Pinheiro GP, Finger M, Nunes MGV, Tagnin SE. The Lacio-Web Project: overview and issues in brazilian portuguese corpora creation. Corpus Linguistics. 2003; 16:14-21.

20 - Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas.1960; 20(1):37-6.

21 - Bashyam V, Taira R. Identifying anatomical phrases in clinical reports by shallow semantic parsing methods. Proceedings of thr IEEE Symposium on Computational Intelligence and Data Mining; 2007 Mar 1 - 2007 Apr 5; Honolulu, H. Washington: IEEE; 2007.

22 - Kailas, A. On medical informatics for pervasive and ubiquitous computing in eHealth. Carolina do Norte: Healthcom; 2012.

23 - Schulz S, Sbrissia E, Nohama P. Cognate mapping - a heuristic strategy for the semi-supervised acquisition of a spanish lexicon from a portuguese seed lexicon. Proceedings of the 20ª International Conference on Computational Linguistics; 2004 Aug 23-27; Geneva, Switzerland. Stroudsburg: Association for Computational Linguistics.

24 - Tretiakov A, Hunter I, Whidett D, Sutinen E. Coding of Medical Records via Restrictive Semantic Topic Tracking. North Melbourne: Health Informatics Society of Australia; 2006.

25 - Hahn U, Wermter J. Tagging medical documents with high accuracy. AI specific application areas - natural language processing. Berlin: Springer; 2004.

Endereço para correspondência/Reprint request to:
Edson José Pacheco
Escola Politécnica
Rua Imaculada Conceição, 1155
Bairro Prado Velho - Curitiba - Paraná
Cep.: 80215-901
E-mail: edsonpacheco@gmail.com

Recebido em: 11-12-2012
Aceito em: 29-5-2013