FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Tecnologias de Codificação Assistida para uma Classificação Internacional de Doenças

Carla Filipa Moura Abreu

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Professor Eugénio Oliveira

Orientadora Fraunhofer: Dra. Liliana Ferreira

19 de Julho de 2013

Tecnologias de Codificação Assistida para uma Classificação Internacional de Doenças

Carla Filipa Moura Abreu

Mestrado Integrado em Engenharia Informática e Computação

Aprovado em provas públicas pelo Júri:

Presidente: Prof. Auxiliar Henrique Cardoso
Arguente: Prof. Auxiliar Daniel Castro Silva
Vogal: Prof. Eugénio Oliveira

---

19 de Julho de 2013

Resumo

Diariamente são produzidos vários volumes de relatórios clínicos nos Hospitais. Grande parte destes documentos são já produzidos no formato eletrónico.

Os dados contidos nos relatórios clínicos são muito úteis, pois servem como base a estudos epidemiológicos e servem para que os Hospitais recebam verba externa [CL95], imperativa ao seu funcionamento. Porém, a estrutura destes documentos não segue nenhum padrão e ao mesmo tempo são escritos de forma livre. Estes dois elementos combinados fazem com que a tarefa de processamento dos documentos seja uma tarefa bastante complexa [CW07].

De forma a ser percetível o conteúdo do documento, entende-se por processamento a forma de extrair do mesmo os dados significantes, como os termos relativos a doenças, sintomas e diagnósticos e a indexação ao mesmo de um ou mais códigos que remetam ao conteúdo nele contido. Neste trabalho, foi utilizado a Classificação Internacional de Doenças na Nona Versão com modificações clínicas (CID-9-CM)[MR09]. Este sistema possui centenas de códigos e um vasto conjunto de regras associados à tarefa de codificação, tornando esta tarefa morosa e complexa [CW07].

Esta tese apresenta um sistema de codificação que visa: numa primeira fase, extrair do texto termos clínicos como nomes de doenças, sintomas e diagnósticos; e numa segunda fase, codificar os termos encontrados de acordo com as regras vigentes.

O sistema de codificação foi projetado em seis componentes: o pré-processamento do documento; a leitura do documento; o processamento de linguagem natural; o reconhecimento de entidades mencionadas; a codificação em CID-9-CM e por fim o processamento do resultado para o entregar ao utilizador. Este sistema foi desenvolvido sobre uma arquitetura de gestão de informação desestruturada (UIMA).

Para a avaliação deste sistema, foram criadas por dois médicos especialistas em cardiologia 22 cartas de alta. Os resultados obtidos apontam para uma precisão de 96.81% na captura de termos clínicos; uma precisão de 90.43% na atribuição do identificador único ao termo clínico. Quanto à codificação final atribuída aos termos clínicos, em 86.17% das vezes o código estava relacionado com a doença.

Palavras-Chave: Mineração de Texto, Extração de Informação, Reconhecimento de Entidades Menciondas, Processamento de Linguagem Natural, Codificação Assistida por Computador, Codificação Automática

Abstract

Currently, several volumes of clinical reports are being produced in the hospitals. Most of these documents are created directly in electronic format.

The data contained in these documents are very important. This data can be used for epidemiology studies and for the hospital receive external funding [CL95]. However, this kind of document do not following any standard and are written in a free-text form. These documents structure make the processing task very hard [CW07].

In order to process the document is essential the relevant terms extraction, these terms correspond to diseases, symptoms and diagnosis names. After this extraction, it is necessary the indexation of codes to terms extracted. In this work, these codes belong to International Classification of Diseases ninth revision with clinical modification (ICD-9-CM) [MR09]. This system has hundreds of codes and has a wide set of rules, these elements make the coding task complex and time-consuming [CW07].

This thesis presents a coding system that aims: the clinical terms extraction from a clinical report, these terms correspond to diseases, symptoms and diagnosis names; and the assigned codes to found terms.

The coding system developed was designed in six components: Document Pre-Processing; Document Reader; Natural Language Processing; Name Entity Recognition; ICD-9-CM Module; Assignment Code Result. This system is developed on top off an unstructured information management architecture (UIMA).

For the evaluation of the system 22 clinical notes were created by two physicians specialists in cardiology. The results that the system got a precision of 96.81% on extraction of clinical terms, a precision of 90.43% in the assignment of unique identifier to clinical term. The assignment of ICD-9-Cm code has a precision of 86.17%.

Keywords: Text Mining, Information Extraction, Name Entity Recognition, Natural Language Processing, Computer Assisted Coding, Automated Coding

Agradecimentos

Ao longo do desenvolvimento deste trabalho recebi o apoio de várias pessoas a quem estou imensamente grata.

O meu primeiro agradecimento vai para os meus orientadores, o Professor Eugénio Oliveira e a Dra. Liliana Ferreira, cujo apoio foi essencial em todas as etapas da realização deste projecto de dissertação.

Gostaria também de agradecer à Instituição Fraunhofer pela oportunidade.

Á Dra. Carla Sousa e ao Dr. Sérgio Leite gostaria de agradecer pelo apoio na criação dos relatórios clínicos imperativos para a avaliação do sistema desenvolvido; ao Dr. Fernando Lopes pelo *feedback* dado relativo ao processo de codificação.

Agradeço também aos meus pais e aos meus irmãos, pela oportunidade que me deram e pela força e coragem que sempre me transmitiram.

Por fim, agradeço aos meus amigos por todos momentos passados e por todo o apoio.

Carla Filipa Moura Abreu

" Informação é poder,
porém se tens tal domínio e não o divulgas,
torna-te responsável pela ignorância alheia."

Ivan Teorilang

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

CA	Codificação Automática
CAC	Codificação Assistida por Computador
CID	Classificação Internacional de Doenças
CID-9	Classificação Internacional de Doenças Nona Revisão
CID-9-MC	Classificação Internacional de Doenças Nona Revisão com Modificações Clínicas
CID-10	Classificação Internacional de Doenças Décima Revisão
CUI	Código Identificador de um termo clínico
EI	Extracção de Informação
HeTop	Portal de Terminologias e Ontologias da Saúde
OMS	Organização Mundial de Saúde
PLN	Processamento Linguagem Natural
REM	Reconhecimento de Entidades Mencionadas
UIMA	Arquitectura de Gestão de Informação Desestruturada
UMLS	Sistema de Linguagem Médica Unificada

Capítulo 1

Introdução

O objectivo deste trabalho de dissertação consiste no desenvolvimento de um sistema capaz de atribuir códigos a relatórios clínicos consoante a informação neles contida.Para isso, o sistema terá que extrair termos clínicos relevantes de um relatório clínico escrito em português e de forma desestruturada. A atribuição de códigos codificação, será realizada segundo a Classificação Internacional de Doenças (CID).

Para que seja possível a tarefa de codificação, torna-se necessário saber quais os termos clínicos relevantes que se encontram no relatório, que neste caso, são termos relativos ao nome de doenças, sintomas ou diagnósticos. Os relatórios clínicos que possuem esta informação contêm os dados escritos de uma forma desestruturada. Para trabalhar com a informação contida nos relatórios clínicos é essencial, numa primeira fase, o desenvolvimento de um módulo de extração de informação. Após esta tarefa é possível passar-se à fase de atribuição de códigos aos termos clínicos detetados, sendo necessário para esta tarefa o desenvolvimento de um módulo de codificação. É imperativo, para o desenvolvimento deste módulo, um conhecimento profundo das regras e restrições inerentes à nomenclatura de classificação.

Este capítulo faculta uma breve descrição sobre o assunto em estudo, sendo também apresentados no mesmo a motivação e os objetivos de trabalho. No final do capítulo é descrita a estrutura da dissertação.

1.1 Enquadramento

Já há muitos anos que, após a realização de uma consulta hospitalar, faz parte do procedimento dos médicos elaborar um relatório que descreva a situação do paciente. Os dados contidos neste documento são muito importantes bem como o seu conhecimento. A necessidade de conhecer estes dados remota do séc XVI, onde surgiu a necessidade de criação de uma classificação de doenças para o desenvolvimento de um estudo estatístico sobre as causas de mortalidade da população. Este estudo intitulado de “London Bills of Mortatity” foi levado a cabo pelo cientista

John Graunt (1620 – 1674). Para o desenvolvimento do mesmo, o pioneiro criou a sua própria classificação de doenças, porém os resultados deste estudo mostraram-se muito vagos, isto devido à classificação criada ou à forma como as causas de morte vinham descritas nos relatórios.

Várias classificações de doenças foram propostas ao longo dos anos, como: a "Genera Morborum" em 1759 proposta pelo cientista Linnaeus (1707 – 1778); a "Nosologia Methodica" desenvolvida por Sauvage (1706 – 1777) e "Synopsis Nosologiae Methodicae" criada por Cullen (1710 – 1790) e publicada em 1785.

Durante o séc. XVIII e felizmente para o progresso da medicina preventiva, a classificação proposta por Cullen foi melhorada por Far (1807 – 1883), permitido uma uniformização da mesma para ser possível o seu uso a nível internacional. Embora o esforço realizado por Far, esta classificação nunca foi aceite a nível internacional. Em 1885, seguindo os conceitos propostos por Far, Bertillan desenvolveu a classificação intitulada "Bertillan Classification of Causes of Death", sendo esta classificação aprovada por vários países.

Os sistemas de classificação começaram a ser desenvolvidos para causas de morte, contudo, Far reconheceu que seria desejável abranger a classificação a outros domínios. Esta consideração foi tida em conta em 1860 no Quarto Congresso Estatístico Internacional.

Atualmente, a classificação adotada internacionalmente é designada de "Classificação Internacional de Doenças (CID)" tendo sido proposta pela Organização Mundial de Saúde (OMS). Esta classificação é utilizada não só para o desenvolvimento de estudos estatísticos e epidemiológicos mas também para facilitar a gestão clínica e de saúde e devido a propósitos económicos. A área económica assume aqui um principal relevo, visto que os hospitais necessitam destes dados para efetuarem os seus pagamentos e para receberem verba externa.

A CID é utilizada para classificar doenças e outros problemas de saúde contidos nos relatórios clínicos, chamando-se a este processo codificação.

É importante mencionar, que hoje em dia continuam a ser produzidos diariamente vários volumes de relatórios clínicos nos hospitais portugueses.

Segundo estudos da Comissão Europeia, 87% das práticas médicas são reportadas computacionalmente. Este valor acresce para 88% em Portugal. O desenvolvimento dos relatórios médicos no formato eletrónico permite-nos pensar na automatização do processo de codificação.

1.2 Classificação Clínica

Atualmente, a codificação clínica é feita com base no sistema de Classificação Internacional de Doenças, a CID que já se encontra na sua décima versão. Porém, no mercado Português ainda não foi adotada esta versão, estando ainda em vigor a nona, ou seja, o CID-9.

O CID foi desenvolvido primeiramente para a codificação de doenças e não de diagnósticos [CL95], o que em termos médicos é bastante diferente, apresentando-se como outro grande desafio para os codificadores [CL95] e para a automatização do processo.

Figura 1.1: Deslocação do Paciente ao Hospital e Escrita do Relatório Clínico

Para a análise dos relatórios clínicos é fundamental utilizar a versão modificada para o efeito, ou seja, o CID-9-MC (Classificação Internacional de Doenças com Modificações Clínicas, nona versão).

1.3 Codificação De Relatórios Médicos

A codificação clínica de relatórios médicos é uma tarefa bastante complexa [CW07]. Esta tarefa requer para o efeito profissionais médicos especializados. Por norma, um médico especializado em codificação clínica trabalha no seu "tempo total de trabalho" com um subconjunto restrito de códigos CID-9-MC de entre uma centena de códigos fornecidos por esta nomenclatura. Acontece, portanto, que várias vezes os codificadores discordam da codificação atribuída pelos colegas [CW07].

É importante referir que, atualmente, o período de tempo entre o momento em que o relatório médico está disponível e o momento em que o código CID-9-MC correspondente lhe é atribuído tende a alargar-se excessivamente.

De uma forma geral todo este processo se divide em duas atividades-chave. A primeira, apresentada pela Figura 1.1 corresponde ao deslocamento do paciente ao hospital e respetiva consulta médica. Enquanto a segunda, representada na Figura 1.2, corresponde à atribuição ao relatório médico do código CID-9-MC correspondente.

1.4 Motivação

A possibilidade de se automatizar o processo de codificação têm vindo a ser estudado desde os anos 90. Para esta automatização tem sido usados princípios de processamento de linguagem natural e codificação assistida por computador. Estas tecnologias permitem a conversão de um

Figura 1.2: Codificação do Relatório Clínico

processo resolúvel manualmente para uma resolução semiautomática. Esta conversão pode trazer ganhos de produtividade a nível de:

- Redução de tempo - o tempo entre a escrita do documento e a codificação.
- Melhoria do processo de codificação.
- Melhoria da informação acessível, podendo esta informação ser utilizada eficazmente e eficientemente para processos de gestão, planeamento, treino e investigações biomédicas.

1.5 Questões de Investigação

As duas grandes questões subjacentes a este trabalho são as seguintes:

- Será possível a extração de termos clínicos relevantes dos relatórios clínicos?
- Será possível automatizar o processo de codificação?

1.6 Objetivos

O âmbito deste projeto é o desenvolvimento de um sistema capaz de analisar relatórios clínicos desenvolvidos no formato electrónico e efetuar a codificação dos mesmos. Neste trabalho em particular, a informação contida nos relatórios clínicos que se pretende codificar está relacionada com os termos relativos aos nomes de doenças, sintomas ou diagnósticos.

O sistema a ser desenvolvido pretende ser um sistema de apoio à decisão, criada para auxiliar o codificador nas suas tarefas.

Para que seja possível o desenvolvimento deste sistema é necessário perceber quais as tecnologias mais aptas para a sua construção, sendo também necessário um vasto estudo da nomenclatura e das regras do CID-9-MC para, perceber como o sistema deve operar.

1.7 Estrutura da Dissertação

O presente documento está organizado da seguinte forma:

Capítulo 2: introduz o estado da arte. Começa por apresentar os assuntos relacionados com o processamento dos relatórios clínicos, ou seja: Mineração de Texto; Extração de Informação; Reconhecimento de Entidades Mencionadas. Seguidamente são apresentadas duas abordagens de codificação; a codificação assistida por computador e a codificação automática.

Capítulo 3: apresenta os recursos e ferramentas principais utilizadas no desenvolvimento do sistema de codificação. Os recursos principais utilizados são os seguintes: Sistema de Linguagem Médica Unificada; Ontologia CID-9-CM; Portal de Terminologias e Ontologias de

Saúde e o MediLexion. Quanto às ferramentas utilizadas, recorreu-se: a uma Arquitectura de Gestão de Informação Desestruturada; o Protégé, o Jena, o Drools Expert Engine e o GraphViz.

Capítulo 4: apresenta a representação de conhecimento desenvolvida e adaptada para este trabalho. No âmbito deste trabalho foi adaptada uma ontologia pré-existente, e foi criado um sistema de regras.

Capítulo 5: apresenta a abordagem utilizada para o desenvolvimento do sistema de codificação. Primeiramente é descrita a arquitectura do sistema, sendo seguidamente apresentadas todas as etapas realizadas.

Capítulo 6: apresenta a avaliação elaborada ao sistema.

Capítulo 7: apresenta as conclusões obtidas neste trabalho.

Capítulo 2

Revisão Bibliográfica

Neste capítulo será abordada a área em que se encontra o trabalho. Primeiramente serão referidos os temas relacionados com o processamento dos relatórios médicos, como o " *Text Mining* ", "Extração de Informação" e o "Reconhecimento de Entidades Mencionadas". Após esta descrição, serão referidas algumas características inerentes à nomenclatura CID-9-MC, como algumas normas impostas ao processo de codificação. De seguida, serão referidas duas formas distintas de se efetuar a codificação recorrendo a *software*: a Codificação Assistida por Computador e a Atribuição Automática de Códigos. Para estas duas diferentes abordagens, as respetivas metodologias existentes no estado da arte serão apresentadas.

2.1 Extração de Conhecimento de Texto

A informação clínica disponível sobre o estado do paciente é reportada de forma livre. O conhecimento relativo ao estado do paciente encontra-se nestes relatórios de forma desestruturada. Para se conseguir perceber computacionalmente a informação contida nestes relatórios é necessário realizar-se a extração da mesma. Cientificamente, está tarefa esta relacionada com a área do *Text Mining*.

O conceito de *Text Mining* é definido por Ah-Hwee Tan como "O processo de extração não trivial de padrões ou conhecimento a partir informação não estruturada" [FS08].

Este é um assunto muito abrangente, como Ah-Hwee Tan refere, "é um assunto multidisciplinar, envolvendo: recuperação de informação, análise de texto, extração de informação, agrupamento, categorização, visualização, tecnologia de base de dados, aprendizagem máquina e *data mining* [FS08].

2.1.1 Extração de Informação

Este trabalho centra-se numa parte específica do *Text Mining*, a Extração de Informação (EI).

"O propósito dos sistemas de Extração de Informação é a extração de informação de domínio específico de textos em linguagem natural" [RJ99], noutras palavras " os sistemas de Extração de Informação possuem um conjunto de padrões de extração que são utilizados de forma a recuperar a informação relevante em cada documento" [Mus99].

Geralmente, a EI permite que sejam elaborados modelos mais ricos e flexíveis através do processamento de um documento [RF07]. Os elementos que podem ser extraídos a partir desta abordagem são os seguintes:

Entidade Corresponde a elementos básicos que podem ser extraídos do texto. Estes correspondem, neste caso, a doenças, sintomas e diagnósticos.

Atributo É uma característica associada à entidade.

Facto Refere-se às relações existentes entre entidades.

Evento O evento é uma atividade ou ocorrência de interesse em que a entidade participa.

Vejamos um exemplo, para a frase "O paciente apresenta pneumonia"podem observar-se os seguintes elementos: entidade - o termo relativo ao nome de uma doença *pneumonia*; atributo - uma característica associada à doença *aguda*.

2.1.2 Text Mining: Na Área da Saúde

O *Text Mining* é utilizado em diversas áreas como: Biologia, Biomedicina [AM06], Saúde, Processamento Sinal, Recuperação de Informação [CW07], entre outras. No presente trabalho focaliza-se na àrea de Saúde.

O *Text Mining* nesta área surgiu recentemente, isto porque os relatórios médicos têm uma particularidade; esta particularidade relacionada com o anonimato dos relatórios médicos, isto porque as informações contidas nos mesmos são confidenciais, respeitando aspetos éticos e sociais [JP08].

Um dos factos específicos da àrea da sáude é que "os relatórios clínicos são uma coleção de relatórios não estruturados sem nenhum requisito na composição" [SWD05]. Estes relatórios, para além de não seguirem nenhum padrão, contêm conceitos médicos passíveis de ser representados de diversas formas . Para a mesma doença podem ser utilizadas abreviaturas, podendo essas abreviaturas variar de clínico para clínico e/ou de instituição para instituição.

Devido às diversas formas que um termo pode assumir, a nível lexical, a terminologia torna-se num grande desafio para o *Text Mining*. Para além disso, os relatórios médicos possuem outros elementos que têm que ser tidos em conta como: caracteres especiais (como hífens), termos com dígitos [KFT98] e sequência de letras sem sentido aparente [JMB11].

A terminologia assume nesta área um papel importante porque permite encontrar relações entre termos médicos. Ao nível do processamento, na parte semântica, "terá que ser criado continuamente neologismos" [AM06] entre os termos médicos.

Para se saber se estamos perante termos médicos (doenças, sintomas ou diagnósticos) é necessário recorrer-se a fontes de conhecimento externo. Para isso, uma das fontes de conhecimento utilizadas em trabalhos nesta área é o Sistema de Linguagem Médica Unificada (UMLS), este sistema contêm uma abundância de nomenclaturas, vocabulários controlados e ontologias [JMB11, Fer11, TR03]. A par do uso do UMLS, poderá ser utilizado também a ontologia GALEN [TR03].

2.1.3 Text Mining e Processamento de Linguagens Naturais

Como foi referido anteriormente, a Extração de Informação é um tópico do *Text Mining*. Como Rajman e Besançon referem " o Processamento de Linguagem Natural pode ser visto como uma ferramenta interessante para o reforço dos procedimentos de extração de informação" [MR97].

O processamento de linguagem natural, "pode ser mais ou menos colocado como quem fez o quê, a quem, quando, onde, como e porquê" [KP07]. Para isto, normalmente, recorre-se a conceitos linguísticos como "partes do discurso"(como nomes, verbos, adjetivos...) e estruturas gramaticais (relação entre termos) [Kum11].

Na área de saúde, em trabalhos onde o objectivo é a extração de termos médicos do texto não estruturado, é frequente a utilização de principios PLN [CW07, Fer11]. Nestes trabalhos, é essencial "perceber o texto escrito, usando conhecimento lexical, sintático e semântico e também informação do mundo real" [Kum11].

A análise semântica no processamento de relatórios médicos é muito importante, isto porque o contexto com que as palavras aparecem na frase influencia o código a atribuir. Expressões como "...não foi diagnosticado...", "... o paciente não evidenciou indícios de ...", "... anteriormente diagnosticado ..." são exemplo de expressões que aparecem associadas a um termo médico que não pode ser codificado [Fer11], apesar de estar presente no texto.

É importante referir que de uma falsa codificação clínica resultam encargos financeiros elevados [JP08].

2.1.4 Reconhecimento de Entidades Mencionadas

Posteriormente à fase de principios PNL resultarão palavras passíveis de corresponder a doenças ou sintomas, terá então que haver uma fase de verificação.

Na fase em que existem palavras passíveis de corresponder a termos médicos é necessário recorrer ao Reconhecimento de Entidades Mencionadas (REM) que consiste em confrontar as palavras com fontes do conhecimento externo, como *sites* anatómicos, a fim de verificar se estas correspondem aos termos pretendidos. Esta análise é importante para verificar se o termo encontrado corresponde realmente a um termo clínico.

2.1.5 Sistema MedInX

O sistema MedInX, corresponde a um sistema desenvolvido para o domínio médico, cujo objetivo é o reconhecimento de termos clínicos em cartas médicas. Este sistema foi desenvolvido

para a língua Portuguesa. Tal como no trabalho a ser desenvolvido, também no presente trabalho, os documentos sobre os quais é realizado o processamento são apresentados de forma não estruturada.

Para o processamento do texto em formato livre são utilizadas técnicas de principios de PLN, para além disto, este sistema também providencia mecanismos de leitura, processamento de dados e utiliza recursos externos.

Esta ferramenta foi construída sobre uma aplicação de gestão de informação desestruturada (UIMA). Sendo a UIMA, uma *framework* que define a arquitetura do sistema.

A nível da língua portuguesa é referido neste trabalho que o número máximo de palavras necessárias para definir uma doença completa são três. Este é um dado de extrema importância para o trabalho a realizar, visto ser esta uma particularidade da língua em questão. Quanto ao nível do processamento do texto, após a sua normalização, são identificados os seguintes elementos:

- Palavras;
- Acrónimos;
- Abreviaturas;
- Números e símbolos de pontuação;
- Delimitadores de frase.

É importante verificar a priori se existem acrónimos no texto, isto porque este elemento pode falsear os delimitadores da frase. A análise dos acrónimos é efetuada sobre uma lista que foi manualmente compilada e possui cerca de 130 acrónimos. Após esta fase, é efetuada a delimitação da frase, seguida da utilização de *part-of-speech (POS) taggin*. O POS *taggin* é responsável por verificar as palavras adjacentes às de um termo para perceber o contexto em que o termo surge.

Para o reconhecimento de entidades é utilizada como fonte de conhecimento o UMLS. Este trabalho obteve uma avaliação de 95 % para a métrica F.

2.2 Classificação Internacional de Doenças

A Classificação Internacional de Doenças (CID) é utilizada para classificar doenças e outros problemas de saúde contidos nos relatórios clínicos. Na nomenclatura CID a um código é associado o termo clínico que o define. Os códigos nesta nomenclatura variam de três a cinco dígitos, e quantos mais dígitos forem utilizados, mais específico é o termo clínico representado. A composição de um código CID-9-MC pode ser observado pela Figura 2.1 , onde a letra X corresponde a um número entre 0 e 9.

Para que seja efectuada a tarefa de codificção, atualmente, um codificador necessita de recorrer ao manual e às regras de codificação, tornando-se esta tarefa complexa e demorada. O manual da CID-9-MC é composto por três volumes:

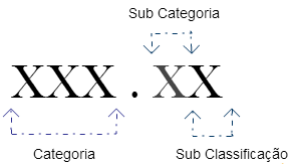

Figura 2.1: Composição de um código CID-9-MC

Volume 1 Índice Alfabético de Diagnósticos

Volume 2 Índice Tabular de Diagnósticos

Volume 3 Índice Alfabético e Tabular de Procedimentos

Para a elaboração de um relatório clínico, o codificador tem que seguir as seguintes etapas:

- Localizar o termo no índice alfabético;
- Procurar notas relacionadas com o termo principal;
- Procurar modificadores associados com o termo principal;
- Procurar sub termos associados com o termo principal;
- Seguir as regras;
- Verificar o número no índice tabular;
- Atribuir o código.

Como foi mencionado anteriormente, a codificação é uma tarefa bastante complexa e demorada, sendo também uma tarefa susceptível de erros, o que torna necessário a automatização da mesma. É de notar que os erros associados ao processo de codificação resultam em encargos financeiros elevados. Os valores dos encargos financeiros resultantes da recuperação dos erros de codificação, em 2007, foi de 25 biliões de dólares nos Estados Unidos.

2.3 Codificação Assistida por Computador

Existem vários problemas com a codificação manual, como a complexidade da tarefa e o vasto conjunto de códigos da nomenclatura ICD-9-CM. Estes problemas fazem com que a tarefa de codificação seja suscetível de erros. Para simplificar este processo de codificação e tornar o resultado da codificação fiável, os investigadores descobriram uma forma de efetuar a codificação de relatórios médicos em formato eletrónico utilizando *software* [JGM05].

As duas formas mais comuns de se realizar a tarefa de codificação através de *software* são a Codificação Assistida por Computador (CAC) [CL95] e a Codificação Automática (CA) [CW07].

O conceito de CAC é definida por Gravin, Watzlaf e Moeini como "o uso de *software* para gerar automaticamente um conjunto de códigos médicos para revisão, validação, usando para isso documentação clínica fornecida pelos profissionais de saúde" [JGM05], para os mesmos autores, o conceito de Codificação Automática é definido como um método que realiza a atribuição de códigos a relatórios médicos sem a intervenção humana [JGM05]. De uma forma geral, a grande diferença entre estes dois métodos é que na CAC é dado um conjunto de códigos para validação, enquanto que na CA é apenas dada a sugestão de um código.

Apesar destas duas formas de efetuar a codificação, é importante referir que a decisão final é sempre do médico codificador. Cabe ao médico codificador validar o resultado final [JGM05].

2.3.1 Codificação Assistida por Computador na Área de Saúde

Os sistemas de codificação assistida por computador na área de saúde ajudam os profissionais de codificação clínica na sua tarefa, mas, a decisão final da codificação clínica é sempre elaborada pelo profissional. É importante referir que no caso da codificação assistida, é dado ao codificador as várias alternativas de codificação para um relatório clínico.

2.3.1.1 Sistema de Codificação em Ambiente Hospitalar

O trabalho intitulado de "Sistema de codificação em Ambiente Hospitalar" [CL95] é um trabalho desenvolvido para a língua francesa, que visa a atribuição de códigos a relatórios médicos. Tal como no trabalho a ser realizado, também neste relatórios médicos a informação se encontra de forma não estruturada.

Numa primeira fase, para o processamento do texto é feita uma análise às características da língua francesa. Da análise às características da língua foi verificado que os médicos tendem a ser concisos no diagnóstico a atribuir. Desta forma, as frases têm em média 5 palavras, de onde se pode concluir que todas elas são importantes para a análise.

Como este sistema opera sobre relatórios médicos escritos sob forma livre, a fase inicial do sistema recorre ao uso de técnicas de principios de PLN. O PLN é usado para a análise morfológica, sintática e semântica. Após esta análise estamos perante termos passíveis de corresponderem a termos clínicos. Neste artigo são focadas duas abordagens para fazer esta verificação. A primeira consiste em fazer a correspondência das palavras passiveis de corresponder a um termo clínico com um dicionário de termos médicos, a segunda verificação é elaborada quando a palavra não está no dicionário. Esta consiste na decomposição das palavras em unidades semânticas baseada em regras. Após a identificação dos termos da frase é obtido o conjunto de palavras mapeadas. Este sistema não deteta só a doença pelo seu nome, mas também por um conjunto de sinónimos que podem estar associados.

Como última etapa é feita a correspondência dos termos clínicos encontrados com o código correspondente na nomenclatura CID-9-MC. A ferramenta desenvolvida foi testada num hospital por clínicos médicos e mostrou um elevado grau de satisfação.

2.3.2 Codificação Autómatica

Como foi referido anteriormente, este sistema opera sem intervenção humana, dado para um relatório médico o código correspondente ao termo presente.

2.3.2.1 Processamento de Linguagens Naturais Clínicas: Auto-Atribuição de Códigos

O trabalho intitulado "Processamento de Linguagens Naturais Clínicas: Auto-Atribuição de Códigos CID-9" [CW07] refere-se a uma ferramenta de codificação de auto atribuição de códigos CID-9 a relatórios médicos.

Quanto à metodologia utilizada para a construção da ferramenta, esta é baseada em princípios de PNL. Esta ferramenta foi desenvolvida em *Python*, sendo utilizado o *Based Natural Language Toolkit* (NLTK). Antes da análise, o texto passa por uma fase de normalização. Passa depois pelos princípios de PLN, para analisar o texto lexical, sintática e semanticamente. São utilizadas técnicas de *POS tagging* para verificar o contexto das palavras na frase. Adicionalmente, são utilizadas expressões regulares para categorizar datas e/ou expressões que apareçam no documento, como por exemplo "há mais de 1 mês".

A nível do módulo CID-9 é criado um dicionário, em Python. Este dicionário contém os códigos e as doenças correspondentes.

Para treinar e testar o sistema, esta ferramenta possui um conjunto de relatórios médicos, os quais foram cedidos pelo "Department of Radiology at the Cincinnati Children's Hospital Medical Center". Estes relatórios têm uma característica particular, estão divididos em duas secções, uma das secções é um campo de texto onde o médico pode descrever a observação clínica do paciente, a segunda é um espaço onde o médico especializado em codificação clínica coloca a lista com os códigos CID-9 correspondentes ao documento em questão.

É importante referir que o conjunto de relatórios médicos utilizados são anónimos, para além desta particularidade, neste caso em específico este conjunto de relatórios possui as seguintes características:

- Duas secções, uma secção com o diagnóstico e outra com a codificação;
- Reais, os relatórios são escritos no âmbito clínico;
- Abrangentes. Representam vários tipos de dados, ou seja, neste conjunto de dados, aparecem doenças muito frequentes, mas também doenças raras.

Os testes realizados a esta ferramenta consistiram na aplicação dos algoritmos de *Naïve Bayes*, *Maximum Entropy*, e *Decision Tree Classifiers*. Após a realização dos testes, a precisão obtida foi de 100 % para a combinação de códigos primários atribuídos comparando com os códigos dos testes de treino.

2.3.2.2 Construção automática baseada em regras de sistemas de codificação em ICD-9-CM

Este artigo [FS08] descreve o Sistema de codificação automático criado. Este sistema baseia-se em regras para a atribuição de códigos a relatórios médicos. Para além do processamento do texto recorrendo a princípios de PLN, esta ferramenta tem em conta algumas características de codificação da nomenclatura CID-9-MC, são elas:

- Um diagnóstico incerto nunca deve ser codificado;
- Um sintoma deve ser omitido quando está incluído no diagnóstico e/ou doença;
- Doenças passadas não devem ser codificadas;
- Tratamentos que não tenham relevância não devem ser codificados.

Este sistema cria regras de inferência que servem como guia à codificação CID-9-MC. As regras criadas permitem excluir sintomas que estão incluídos na doença. A atribuição de *labels* e as funcionalidades do sistema inicial baseado em regras foram treinados pelo algoritmo de árvores de decisão c4.5.

Este sistema de regras permitiu que a ferramenta de codificação aumentasse o seu desempenho. Com este, houve um aumento de desempenho de 1.5%. A nível de avaliação, a métrica F1 possui um valor de 85.57% para o conjunto de treino.

Este sistema a par das regras da nomenclatura CID-9-MC tem a possibilidade de lidar com o uso de sinónimos para a descoberta do diagnóstico, porque:

- instituições usam nomenclaturas próprias;
- médicos usam abreviaturas não padronizadas.

Esta variante possibilita o aumento de léxico do sistema e diminui o erro de codificação em 30%, no entanto, os autores referem que não existe um sistema de codificação capaz de lidar com todas as formas com que pode ser representado um determinado termo.

2.4 Resumo e Conclusões

Em todos os trabalhos estudados para a análise de relatórios clínicos é evidenciado o uso de principios de PLN.

Um dos aspetos principais a ter em conta são as características da língua para a qual a ferramenta se destina. Cada língua possui características específicas que são importantes para o processamento do texto. Outro elemento que tem especial relevo nesta área é a múltipla forma com que um termo clínico pode ser representado. Como foi verificado em estudos realizados, ao ser aumentado o léxico, considerando mais formas de representação de termos clínicos reduz-se o erro de codificação. Tal como os autores destacam, apesar de efetivamente se diminuir o erro com

o aumento de léxico, é impossível prever todas as formas que um termo pode ser representado [FS08].

Quanto à codificação clínica, como já foi referido, é uma tarefa complexa e morosa. Para esta tarefa é necessário o entendimento da nomenclatura CID-9-MC, bem como das suas restrições.

Para se realizar esta tarefa eletronicamente pode recorrer-se ao desenvolvimento de um sistema de Codificação Assistida por Computador ou por um sistema de Codificação Automática. Em ambos os casos, é importante referir que quem avalia a viabilidade e a performance do sistema são os profissionais clínicos especializados em codificação [JGM05].

O sistema a ser desenvolvido no âmbito deste trabalho consistirá num sistema de apoio à decisão, sendo uma aplicação de Codificação Assistida por Computador.

Tal como nos trabalhos realizados para o mesmo fim, pode concluir-se que é necessário analisar o documento a fim de serem encontradas especificidades:

- A nível da língua em que está escrito;
- Da organização a que se destina (neste caso, se utiliza uma nomenclatura própria para definir os termos médicos).

Para o reconhecimento de termos médicos, será adotada uma abordagem de Reconhecimento de Entidades Mencionadas. Nesta fase, os termos serão confrontados com uma fonte de conhecimento externo - o Sistema Unificado de Linguagem Médica, será também criado ou utilizado um dicionário com as abreviaturas mais comuns contidas nos relatórios médicos.

Paralelamente, será construído o módulo CID-9-MC, contendo os códigos correspondentes aos sintomas, doenças e diagnósticos. Na atribuição dos códigos no relatório clínico terá que ser tida em conta a hierarquia desta nomenclatura. Segundo as regras de codificação, o sistema global deverá incluir as seguintes funcionalidades:

- Detenção de negações e palavras de incerteza - isto porque um diagnóstico incerto nunca pode ser codificado;
- Detenção de palavras que se remetam ao passado - tratamentos ou doenças que não tenham relação directa com o diagnóstico atual não podem ser codificados;
- Sintomas incluídos numa doença presente no relatório deve ser excluídos.

Foram apresentadas duas abordagens distintas de codificação, em ambas o resultado final depende do codificador. A grande diferença entre as duas abordagens é que na codificação assistida por computador, é dada uma lista de possibilidades ao codificador, enquanto que na codificação automática é só mostrada uma opção.

O resultado para o trabalho de codificação assistida por computador foi subjectivo, pois teve por base a opinião de um codificador. Para a outra abordagem os resultados podem ser visualizados através da Tabela 2.1. Porém, é necessário ter em conta que o sucesso dos métodos estatísticos e baseados em aprendizagem dependem do largamente volume de relatórios anotados que tem para treinar o sistema.

Reference	Method	F1
[JPW06]	Modelo Máxima Entropia and Support Vector Machine	88.2 %
[FS08]	Regras Inferência	85.57 %
[CW07]	Métodos Estatísticos	77.0 %
[IGU07]	Lucene	66.9 %
[IGU07]	BoosTexter	mais do que Lucene
[IGU07]	Codificação Baseada em Regras	mais do que BoosTexter

Tabela 2.1: Resultados obtidos de diversas abordagens de codificação

Capítulo 3

Recursos e Ferramentas

Neste capítulo são descritos os recursos e as ferramentas principais utilizadas no desenvolvimento deste sistema.

Inicialmente na secção 3.1 são apresentados os recursos utilizados, sendo eles: o Sistema de Linguagem Médica Unificado, utilizado para a verificação dos termos clínicos; a ontologia CID-9-MC, utilizada no processo de codificação; o Portal de Ontologias e Terminologias da Saúde e o MediLexiom, utilizados na avaliação do sistema.

Na secção 3.2 é apresentada uma breve descrição das ferramentas de processamento utilizadas no desenvolvimento deste sistema e a sua utilidade. De uma forma genérica foram utilizadas infra-estruturas de processamento de linguagem natural para o desenvolvimento do módulo de extração de informação; uma ferramenta para a gestão da ontologia; regras para o desenvolvimento das regras específicas da classificação e uma ferramenta para o desenvolvimento de representações gráficas.

3.1 Recursos

Para o desenvolvimento do sistema foi necessária a utilização de fontes de conhecimento externo: o sistema de linguagem médica unificado e a ontologia CID-9-MC. Estes dois recursos foram usados em módulos distintos, o primeiro foi utilizado no módulo de extração de informação de forma a perceber se os termos contidos no relatório clínico eram realmente termos clínicos; e o segundo foi utilizado no módulo responsável pela codificação, de forma a dotar o sistema desse conhecimento.

Para a realização da avaliação foram utilizados dois portais: o Portal de Terminologias e Ontologias de Saúde e o MediLexion. O primeiro com o objetivo de verificar se os termos que estavam a ser extraídos do relatório eram realmente termos clínicos; e o segundo de forma a verificar se a codificação atribuída fazia ou não sentido.

3.1.1 Sistema de Linguagem Médica Unificado

Em 1836 foi fundada a Biblioteca Nacional de Medicina dos Estados Unidos [oM12]. Em 1986 esta instituição começou um longo estudo e desenvolvimento do sistema de Linguagem Médica Unificada.

Prevendo um crescimento da informação biomédica disponível de forma eletrónica esta instituição encontrou a necessidade de construção de um sistema que facilitasse o desenvolvimento de sistemas de informação avançados capazes de recuperar e integrar informação proveniente de diversas fontes, como bases de dados biográficas; sistemas de relatórios clínicos e bases de conhecimento [TPSJNH06].

A maior barreira encontrada aquando do desenvolvimento deste sistema está relacionada com o chamado *"Naming Problem"*. Este problema consiste na recuperação e integração de informação derivadas de fontes distintas, isto acontece porque o mesmo termo pode ser descrito de várias formas. Atualmente, o UMLS é uma colecção de vocabulário controlado na área da medicina. Este sistema fornece três grandes fontes de conhecimento:

Metathesaurus: coleção de termos médicos e suas relações;

Rede semântica: agregação de relações e categorias utilizadas para classificar os termos no Metathesaurus;

Léxico especialista: base de dados com informação sintáctica, morfológica e ortográfica para a área biomédica.

Na realização deste trabalho foi utilizada a versão de 2011. Nesta versão o sistema tem a capacidade de identificar 77 112 conceitos diferentes e 157 675 conceitos de nomes para a língua portuguesa.

De forma a utilizar-se este recurso, foi adoptada a metodologia adotada pelo sistema PharmInX [Agu12], na qual foi desenvolvido uma base de dados em SQL com os dados relevantes fornecidos pelo sistema.

Neste trabalho em particular pretende-se conhecer os termos relativos a nomes de doenças, sintomas e diagnósticos, o grupo semântico do UMLS que nos providencia esta informação é o Grupo das Doenças, cujas componentes podem ser observadas na Figura 3.1.

Os termos clínicos na base de dados possuem as seguintes características: definição do termo clínico; identificador do termo clínico (CUI); e o grupo a que o termo clínico pertence. Por exemplo, o termo clínico Miocardite: o seu descritor é miocardite; o seu identificador é o C0027059 e pertence ao grupo de Doença ou Síndrome.

3.1.2 Ontologia CID-9-MC

A ontologia utilizada neste sistema foi criada no âmbito do sistema MedInX [Fer11]. Esta ontologia é uma formulação em OWL (linguagem de definição de ontologias) da classificação

Grupo Doenças

> Anormalidades Adquiridas;
> Anormalidades Anatómicas;
> Bactérias;
> Disfunção Molecular ou celular;
> Doenças ou Síndromes ;
> Descobertas;
> Ferimento ou Envenenamento;
> Disfunção Mental ou Comportamental;
> Processo Neoplástico;
> Função Patológica;
> Sinais ou Sintomas.

Figura 3.1: Componentes do Grupo de Doenças

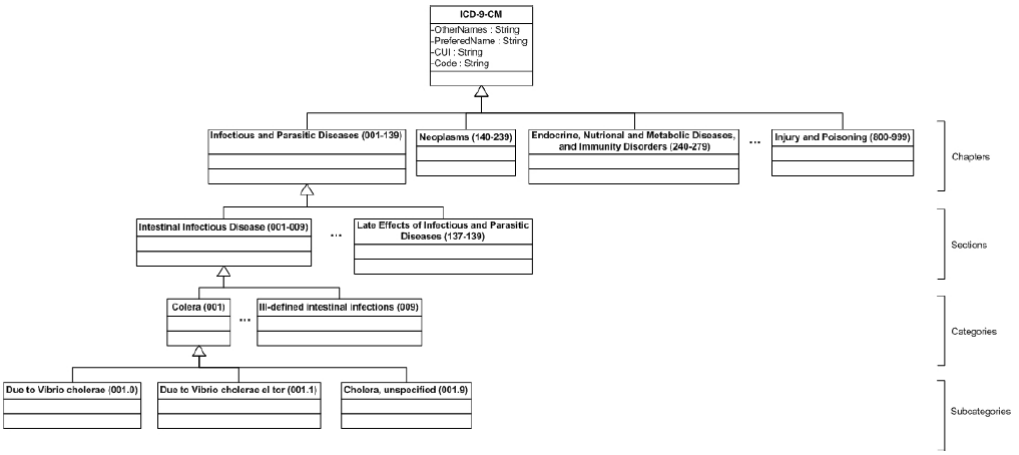

Figura 3.2: Organização da Ontologia

Capítulo	Descrição	Códigos
1	Doenças Infecciosas e Parasitárias	001 -139
2	Neoplasmas	140 - 239
3	Doenças das glândulas endócrinas, da nutrição e do metabolismo	240 - 279
4	Doenças do sangue e dos órgãos hematopoéticos	280 - 289
5	Transtornos mentais	290 - 319
6	Doenças do sistema nervoso e dos órgãos dos sentidos	320 - 389
7	Doenças do aparelho circulatório	390 - 459
8	Doenças do aparelho respiratório	460 - 519
9	Doenças do aparelho digestivo	520 - 579
10	Doenças do aparelho geniturinário	580 - 629
11	Complicações da gravidez, do parto e do puerpério	630 - 679
12	Doenças da pele e do tecido celular subcutâneo	680 - 709
13	Doenças do sistema osteomuscular e do tecido conjuntivo	710 - 739
14	Anomalias congênitas	740 - 759
15	Algumas afecções originadas no período perinatal	760 - 779
16	Sintomas, sinais e afecções mal definidas	780 - 799
17	Lesões e envenenamentos	800 - 999

Tabela 3.1: Capítulos da nomenclatura CID-9-CM

internacional de doenças. Neste trabalho a ontologia foi utilizada no módulo responsável pela codificação de forma a dotar o sistema dessa informação.

Nesta classificação, os códigos estão organizados hierarquicamente como ilustra a Figura 3.2. Esta hierarquia pode ser apresentada em quatro níveis:

17 Capítulos: onde as doenças são agrupadas de acordo com a sua categoria principal (por exemplo, na Tabela 3.1, o primeiro capítulo designado "Doenças infecciosas e parasitárias" representa a categoria, este capítulo engloba todas as doenças desta categoria, estas situam-se entre os códigos 001 e o 139; a Figura 3.3 representação em Protégé os capítulos do CID-9-MC);

Os Capítulos têm secções: o conjunto de códigos de três dígitos representa uma doença simples ou um conjunto de condições semelhantes (por exemplo, códigos entre 001 a 009 – representam todas as doenças relativas a infeções intestinais);

As Secções têm categorias: código com três dígitos representa uma doença simples (por exemplo, o código 001 representa a doença Cólera);

As Categorias têm subcategorias: a subcategoria de um código é representada pelo seu quarto dígito. Este dígito adicional especifica a informação relacionada com a doença.

Na ontologia utilizada, a ontologia CID-9-MC, cada código é representado por uma classe. Cada classe tem os seguintes atributos:

Nome (Preferred Name): é o nome utilizado para descrever a doença;

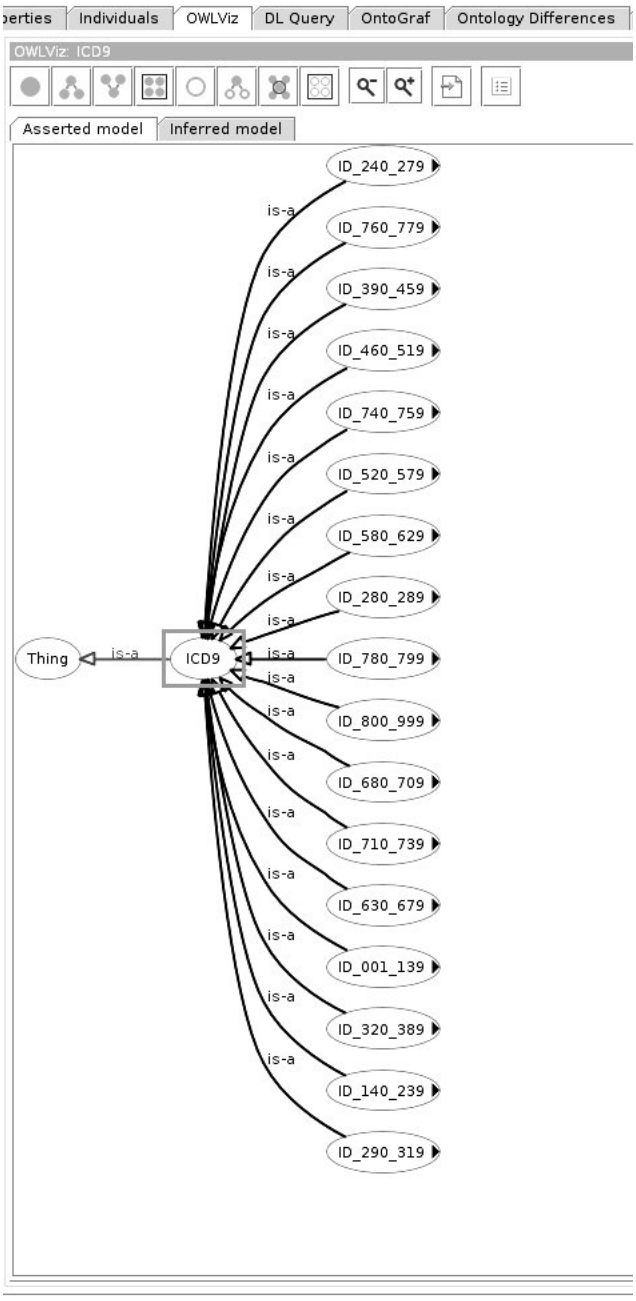

Figura 3.3: Representação em Protégé dos Capítulos do CID-9-MC

Outros Nomes (Other Names): corresponde a outros nomes que a doença pode tomar (ou seja, sinónimos);

Identificador da doença (CUI): corresponde ao identificador único do conceito clínico segundo o UMLS;

Código da Classificação Internacional de Doenças (Code): corresponde ao código ICD-9-CM relativo ao termo clínico.

Por exemplo, o código CID-9-MC 429.0 representa Miocardite, este elemento possui as seguintes características: nome, *Miocardite*; outros nomes: *Miocardite N.E.* (não especificada); identificador da doença, *C0027059* e como código da Classificação Internacional de Doenças o código *429.0*.

Como foi referido a ontologia foi desenvolvida no sistema MedInX, a mesma foi desenvolvida programaticamente utilizando a API da Jena e instanciando automaticamente os recursos do UMLS.

3.1.3 Portal de Terminologias e Ontologias da Saúde

O portal de terminologias e ontologias da saúde (HeTop) foi proposto por *"Chu de Rouen*" com a colaboração do *"I'INSA de Rouen"*, *LERTIN* e a sociedade *MONDECA*.

Como o nome indica, o HeTop, é um portal que fornece terminologias e ontologias de saúde como: *"Biologie Hors Nomenclature"*; Classificação Comum dos atos médicos; CID-9; CID-10; Rubricas de assuntos médicos (MeSH – Medical Subjects Headings) e a nomenclatura sistematizada de medicina (SOMED).

Este portal foi utilizado para a avaliação do sistema desenvolvido, de forma a verificar se os termos extraídos pela plataforma eram realmente termos clínicos e se estavam a ser bem atribuídos os identificadores.

3.1.4 MediLexicon

O *MediLexicon* é um *website* que contém pesquisas médicas, notícias e recursos médicos e farmacêuticos.

Este *website* tem vários recursos à disposição, como: uma lista de abreviaturas médicas (mais de 200 000); dicionários médicos; noticias médicas; informação sobre equipamento médico e instrumentos cirúrgicos e permite ainda a pesquisa de medicamentos, códigos CID-9-MC entre outras pesquisas médicas.

Nesta dissertação, este *website* foi utilizado para a avaliação do sistema para se verificar se a codificação estava a ser realizada de forma correta.

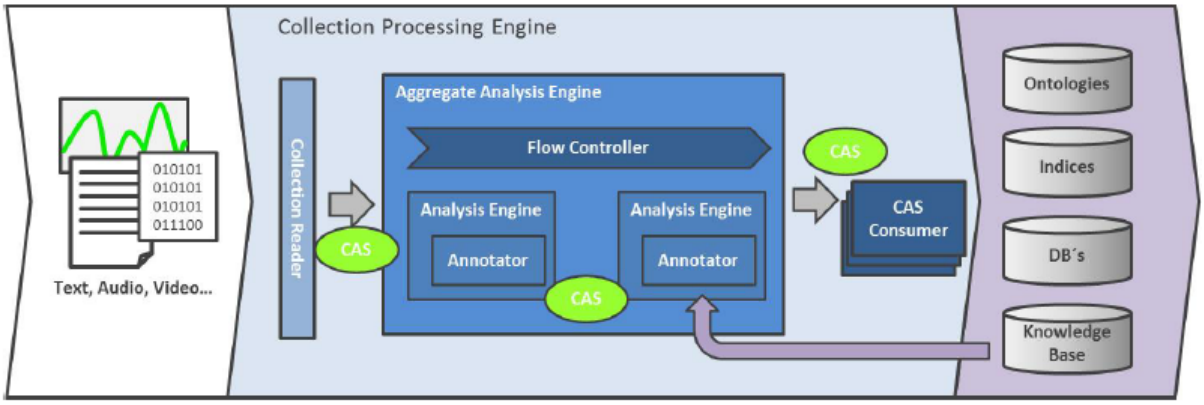

Figura 3.4: Arquitectura do UIMA (figura original [Fer11])

3.2 Ferramentas

Uma importante tarefa neste trabalho está relacionada com a escolha das ferramentas de processamento. Foram vários os aspectos considerados para a escolha das mesmas, sendo um dos aspectos primordiais a possibilidade de integração entre elas.

Neste secção serão apresentadas as ferramentas escolhidas para o desenvolvimento do sistema.

3.2.1 Arquitetura de Gestão de Informação Desestruturada

A arquitetura de gestão de informação desestruturada (UIMA) é uma *framework* extensível e escalável. Esta *framework* é utilizada exclusivamente para aplicações de processamento de linguagem natural [KR08], permitindo a criação, integração e implementação de soluções de gestão de informação desestruturada [FL04].

A UIMA foi desenvolvida pela IBM, estando agora incubada na *Apache Software Foundation* [oM12] onde está disponível como um projeto *open source*.

Esta *framework* compreende duas fases: a fase de análise e a de entrega. Na primeira, os documentos são armazenados e analisados e os resultados são produzidos e armazenados de acordo com os requisitos da fase de entrega. Como foi referido o UIMA visa trabalhar com informação desestruturada. Sobre o documento de entrada, que neste caso em particular se refere a relatórios clínicos, a fase de entrega pode englobar a descoberta de tokens (termos individuais), detenção da classe semântica do termo, podendo também fazer uso de recursos externos como terminologias e ontologias para realizar a anotação das entidades mencionadas. A fase de entrega é responsável por entregar o resultado ao utilizador e isto é realizado através de uma interface da própria *framework*.

As principais funcionalidades da UIMA são o tipo de sistema (UIMA Type System) e a estrutura comum de análise (CAS). O tipo de sistema corresponde a uma definição declarativa de um objeto [KVH09], é nesta componente em que os anotadores são definidos e as anotações realizadas armazenadas. A CAS é responsável pelo armazenamento da informação desestruturada ajudando também na realização das anotações.

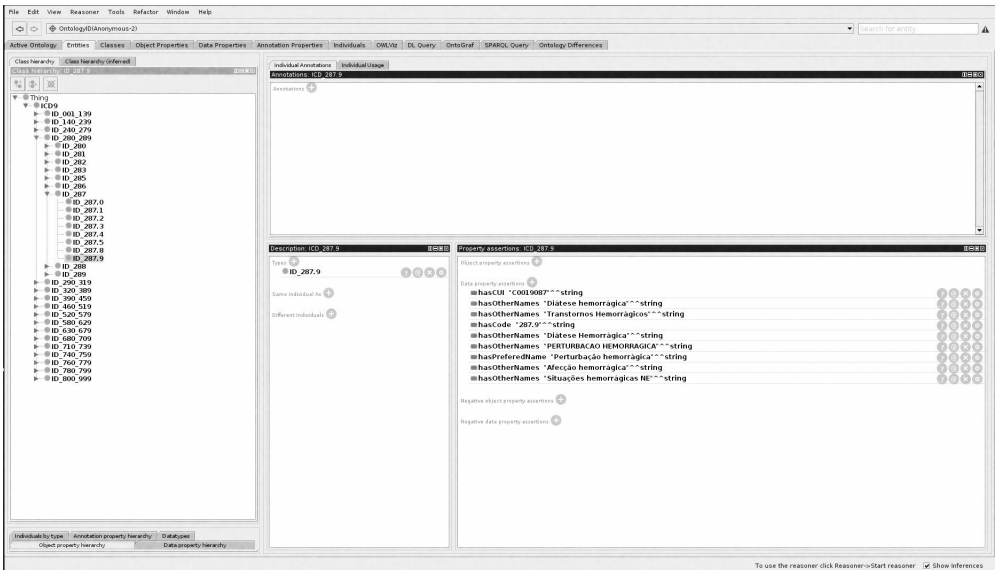

Figura 3.5: Interface do Protégé

O CAS providencia mecanismos de análise, como acesso de leitura para saber o conteúdo de análise e de leitura e escrita, para aceder ao conteúdo a analisar e escrever o resultado da análise. De uma forma geral, é no CAS que ficam armazenados os dados para análise bem como os resultados das anotações, é a partir deste que se conseguem recuperar anotações realizadas previamente ao documento [Zia]. Para que seja possível essa tarefa, este mecanismo possui duas componentes essenciais: o Leitor (Collection Reader) e o consumidor (CAS Consumer). O primeiro é responsável por adquirir componentes e inicializar o sistema para análise. A segunda componente é responsável pelo processamento final.

O fluxo deste recurso pode ser observado pela Figura 3.4.

3.2.2 Protégé

A ferramenta escolhida para se trabalhar com a ontologia CID-9-MC foi o Protégé-OWL, sendo esta uma extensão da ferramenta original do Protégé [Pro13]. Esta ferramenta foi desenvolvida pelo Centro de Investigação de Informação Médica de Stanford pela Escola de Medicina da Universidade de Standord. Inicialmente o Protégé foi desenvolvido como uma pequena aplicação desenhada para o domínio médico [JHGT02].

O Protégé é um editor de ontologias e uma base de conhecimento *open-source*. Esta framework é caracterizada por uma arquitectura aberta, extensível e customizável. Esta framework permite a criação e edição de ontologias a partir de uma interface intuitiva. A interface deste editor pode ser observada na Figura 3.5. O Pótégé é escrito em Java e pode correr sobre variados sistemas operativos.

A vantagem desta *framework* é a representação da linguagem ser independente o que não acontece com as seguintes ferramentas: *WebOnto* [Dom98], *OntoSaurus* [BSR97] e *Ontolingua*

2011 Addenda

The addenda below are effective October 1, 2011.

Note: These addenda were created on 06/10/2011.

- ICD-9-CM Index Addenda [PDF - 252 KB]. 43 pp. (10/01/2011)
- ICD-9-CM Tabular Addenda [PDF - 182 KB]. 49 pp. (10/01/2011)

Note: These errata were posted on 06/30/2011.

- ICD-9-CM Index Addenda Errata [PDF - 16 KB]. 1 pp.
- ICD-9-CM Tabular Addenda Errata [PDF - 32 KB]. 2 pp.

2010 Addenda

The addenda below are effective October 1, 2010.

Note*: These addenda were posted on 05/20/2010.*

- ICD-9-CM Index Addenda [PDF - 155 KB]. 26 pp. (10/01/2010)
- ICD-9-CM Tabular Addenda [PDF - 132 KB]. 39 pp. (10/01/2010)

Figura 3.6: CID-9-MC Adendas

Server [AFR97].

3.2.3 Jena

A *framework* Jena foi desenvolvida em 2000 pelo programa de investigação web semântica dos Laboratórios HP [Jen13]. Em 2010, o Jena foi adoptado pela Fundação Apache Software.

O Jena é uma *framework* em Java para desenvolvimento de aplicações de *Web* Semântica baseada no Consórcio *Word Wide Web* (W3C) para a descrição de recursos (RDF) e para a linguagem de ontologia web (OWL). Esta é uma *framework open source* e permite a programação nos seguintes ambientes RDF; Esquema RDF e OWL.

No sistema de codificação desenvolvido no âmbito deste trabalho, o Jena foi utilizado para se aceder ao conteúdo da ontologia CID-9-MC.

3.2.4 Java Drools Engine

De acordo com as regras de codificação existem dois géneros de regras, as regras gerais e as regras específicas. As regras gerais são aplicáveis sempre, enquanto as regras específicas são apenas aplicáveis aquando da presença de determinados elementos. Para o desenvolvimento das regras específicas, foi utilizada a ferramenta Drools [Dro13]. Esta ferramenta encontra-se disponível na comunidade JBoss, sendo orientada a objetos e apta para a linguagem Java.

A lógica subjacente ao Drools baseia-se no algoritmo RETE. Esta plataforma permite que os utilizadores criem regras muito próximas da linguagem natural. As maiores vantagens desta ferramenta são as seguintes:

Programação Declarativa: usando as regras pode-se de uma forma simples expressar a solução para um problema e testa-lo;

Separação da Lógica e dos dados: os dados estão associados aos objetos e a lógica está associada às regras;

Rapidez e escalabilidade: a lógica subjacente ao Drools (algoritmos RETE e Leaps) providencia uma forma eficiente de combinar as regras com os dados dos objetos;

Centralização do conhecimento: usando regras é possível criar um repositório de conhecimento executável. Idealmente, estas mesmas regras poderão ser utilizadas como documentação.

Integração Ferramenta: esta ferramenta pode ser integrada com o eclipse;

Percepção das regras: o programador pode desenvolver as regras próximas da linguagem natural, sendo possível pessoas sem conhecimentos técnicos adicionarem, modificarem e removerem regras de uma forma fácil.

O Drools foi escolhido pelas seguintes razões pela complexidade intrínseca à codificação e pelo aparecimento de várias adendas à classificação (como se pode visualizar na Figura 3.6). Esta ferramenta permite que sejam feitas alterações às regras sem que seja necessária a recompilação do sistema, conseguindo desta forma elaborar alterações de uma forma simples e eficaz.

3.2.5 GraphViz

De forma a transmitir ao codificador todas as etapas percorridas pelo sistema até à escolha do código final foi necessário a criação de um gráfico. Para tal, foi utilizada uma ferramenta *open-souce* de visualização de grafos, o GraphViz [Gra13]. Esta ferramenta proporciona uma forma de representar a informação em forma de grafos, diagramas e redes. O GraphViz possui várias funcionalidades para a criação de gráficos, como opções de cores, fontes, tipologias para os nós, vários estilos de ligações entre outros elementos; permite também a criação de grafos em várias linguagens, como: *dot*, *neato*, *fdp*, *sfdp*, *twopi*, e *circo*.

A possibilidade de integração desta ferramenta com o eclipse fundamenta a escolha desta ferramenta. De forma a mostrar as diferentes etapas percorridas pelo sistema no processo de codificação foi utilizada a linguagem dot, isto porque esta linguagem permite a criação de grafos hierárquicos e direcionados.

Capítulo 4

Representação do Conhecimento

Para o desenvolvimento deste sistema foi necessário a utilização de fontes de conhecimento externo. Neste trabalho as duas fontes de conhecimento utilizadas foram uma ontologia CID-9-MC e um sistema de regras.

A ontologia utilizada, como foi referido no capítulo anterior foi desenvolvida noutro sistema. Porém após o estudo das regras de codificação do sistema de classificação utilizado verificou-se a necessidade de introdução de novos campos. O sistema de regras utilizado, foi criado de raiz neste projeto, tendo em conta que por vezes podem surgir algumas modificações que devem ser facilmente introduzidas no sistema para garantir que o mesmo se mantenha facilmente atualizado.

Este capítulo apresenta estas duas representações de conhecimento.

4.1 Ontologia CID-9-CM

Neste trabalho foi utilizada a ontologia apresentada na secção 3.1. Sendo esta ontologia uma representação hierárquica dos códigos da nomenclatura CID-9-MC. Após o estudo das regras de codificação inerentes ao CID-9-MC foi verificada a necessidade de introdução de alguns atributos que ainda não estavam presentes na ontologia, como:

See (Ver): corresponde a um código CID-9-MC;

SeeAlso (Ver Também): corresponde a um código CID-9-MC;

Symptoms (Lista de Sintomas): corresponde a uma lista de códigos.

De acordo com a nomenclatura, o sistema deve ter as seguintes características: See, que indica que outro código tem que ser obrigatoriamente atribuído ao documento; o SeeAlso, que indica que o codificador terá que analisar se no contexto do documento faz sentido a introdução de um código adicional e Symptoms, porque segundo as regras é importante conhecer os sintomas associados com uma doença em específico, pois, se no mesmo documento estiver presente uma doença e

um sintoma a ela associado, o código do sintoma tem que ser excluído. A adição destes novos atributos à ontologia foi realizada utilizando o Protégé.

4.2 Regras

O processo de atribuição de códigos CID-9-MC aos relatórios clínicos é uma tarefa bastante complexa e demorada. Para a realização desta tarefa, o codificador tem que ter um conhecimento amplo da nomenclatura e do sistema de regras associados.

Após o estudo das regras de codificação, foi verificada a existência de dois géneros de regras: as regras gerais e as regras específicas. As regras gerais estão sempre associadas ao processo de codificação e tem que fazer parte do processo de codificação. Quanto às regras específicas, estas somente se evidenciam aquando do aparecimento de situações particulares.

Alguns exemplos de regras gerais são:

- Nível de detalhe da codificação: os códigos relativos a procedimentos e diagnósticos devem ser sempre usados no maior número de dígitos possíveis;
- Condições que fazem parte integrante da doença: sinais e sintomas que estão associados com a doença não devem ser codificados;
- Condições que não fazem parte integrante da doença: sinais e sintomas que aparecem no relatório clínico e não fazem parte das doenças neles presentes devem ser codificados [ICD10].

Como estas regras abrangem todo o processo de codificação, estas regras fazem parte do algoritmo de codificação. Porém, para as regras específicas que só se evidenciam após o aparecimento de uma situação particular, foi criado um sistema de regras utilizando o Drools Engine (mencionado no capítulo 3).

O sistema de regras elaborado, foi projetado para o capítulo da cardiologia, isto porque o projecto contou com o apoio de dois médicos especialistas que criaram os relatórios médicos para a avaliação do sistema, nesta área também. A criação deste sistema é muito útil, pois permite a alteração, remoção ou edição de regras sem ser necessário a recompilação do sistema. Como se pode visualizar pela Figura 3.6, por vezes surgem certas adendas que mudam o código a atribuir, com este sistema de regras consegue-se facilmente introduzir estas alterações, permitindo a consistência e manutenção do sistema.

Para o desenvolvimento deste sistema foram desenvolvidos os seguintes métodos:

setCode: permite que seja realizada a alteração de um código;

getCode: permite receber o código presente;

CuiInGroup: verifica se um determinado identificador do termo clínico está presente no relatório;

haveCodes: verifica se um conjunto de códigos CID-9-MC estão presentes no documento;

createRelationship: permite a criação de uma relação;

setRules: permite que a regra seja alterada;

setAdditionalCode: permite que seja alterado o código adicional;

setSecundaryCode: permite que o código secundário seja modificado;

setSendToPresentList: permite que o código seja movido para a lista de códigos presentes da lista de códigos excluídos.

Para a construção das regras, foram utilizados os métodos mencionados. A estrutura das regras pode ser definida da seguinte forma:

```
Rule "Rule Name"
When
"Rule Condition"
Then
 "Behaviour expected when the rule condition are checked"
end
```

4.3 Resumo

O sucesso da tarefa de codificação está diretamente relacionado com o conhecimento da nomenclatura, regras e propriedades do CID-9-MC. Neste capítulo, foi apresentada a metodologia utilizada para a modificação da ontologia, bem como para o desenvolvimento do sistema de regras específicos.

O sistema de regras foi desenvolvido de forma a ser facilmente extensível e atualizável.

Capítulo 5

Sistema de Codificação

O sistema desenvolvido neste trabalho intitula-se de *Assisted Coding Technologies For International Classification of Diseases*. O objetivo deste trabalho corresponde, numa primeira fase, ao desenvolvimento de uma componente de extração de informação capaz de reconhecer os termos clínicos presentes no relatório e as suas características. Seguidamente, utilizando a informação obtida será possível a realização da tarefa da codificação, sendo necessário para tal, o desenvolvimento de um módulo de codificação.

Como já foi referido este trabalho destina-se ao domínio clínico e a codificação realizar-se-á para a nomenclatura CID-9-MC.

A arquitetura do sistema foi desenhada de forma a garantir os seguintes requisitos: fácil atualização, flexibilidade e consistência.

Neste capítulo será descrita a arquitetura bem como a metodologia utilizada para o desenvolvimento do sistema de codificação.

5.1 Arquitetura

O sistema desenvolvido tem duas componentes essenciais, sendo que a primeira corresponde à extração dos termos clínicos a partir de relatórios clínicos e a segunda corresponde á atribuição da classificação segundo a nomenclatura CID-9-MC.

Para o desenvolvimento da primeira componente foram utilizados princípios de processamento de linguagem natural (PLN). O objetivo desta componente é a extração das diferentes entidades e das suas relações de um relatório escrito de forma livre na língua portuguesa. Para o desenvolvimento desta tarefa foram utilizadas fontes de conhecimento externo como ontologias e terminologias.

A segunda componente necessita do resultado da primeira componente, devido a esse motivo estas componentes são realizadas sequencialmente. O objetivo da segunda componente é a atribuição de códigos, sendo esta etapa dividida em duas sub-etapas, que passa pelo cumprimento das regras gerais seguido do cumprimento das regras específicas.

A arquitetura deste sistema pode ser apresentada em seis fases distintas:

- Pré-Processamento do documento;
- Leitor do documento;
- Princípios de Processamento de Linguagem Natural;
- Reconhecimento de Entidades Mencionadas;
- ICD-9-CM;
- Resultado da Codificação.

5.2 Pré-Processamento do Documento

Os relatórios clínicos por vezes encontram-se em formatos distintos, para que seja possível trabalhar com a informação contida nos mesmos é necessário a normalização dos mesmos. Nesta fase de pré-processamento do documento, o documento é convertido do seu formato original (word ou pdf) para um ficheiro XML. É importante referir que este é o formato mais indicado a utilizar pois permite que se façam anotações no relatório a partir de *tags*, sendo assim possível trabalhar com informação não estruturada.

5.3 Leitor do Documento

Após a fase de pré-processamento segue-se a fase do leitor do documento, sendo esta etapa responsável pela inicialização do sistema. Nesta fase são lidos os relatórios a analisar e carregadas as fontes de conhecimento, sendo esta informação enviada para a estrutura comum de análise, para futuro processamento.

Também os diferentes tipos de anotadores são gravados na estrutura comum de análise. As anotações correspondem a entidades ou relações extraídas consoante a anotação que se pretende realizar.

5.4 Princípios de Processamento de Linguagem Natural

De forma a extrair o conhecimento básico do relatório clínico são utilizados princípios PLN.

O objetivo desta fase é a identificação de elementos simples do texto, como: *tokens*, ou seja, visa a identificação de palavras, abreviaturas e outros elementos simples; frases, pretende reconhecer onde se inicia e onde termina uma frase; *stop-words*, que corresponde à identificação de

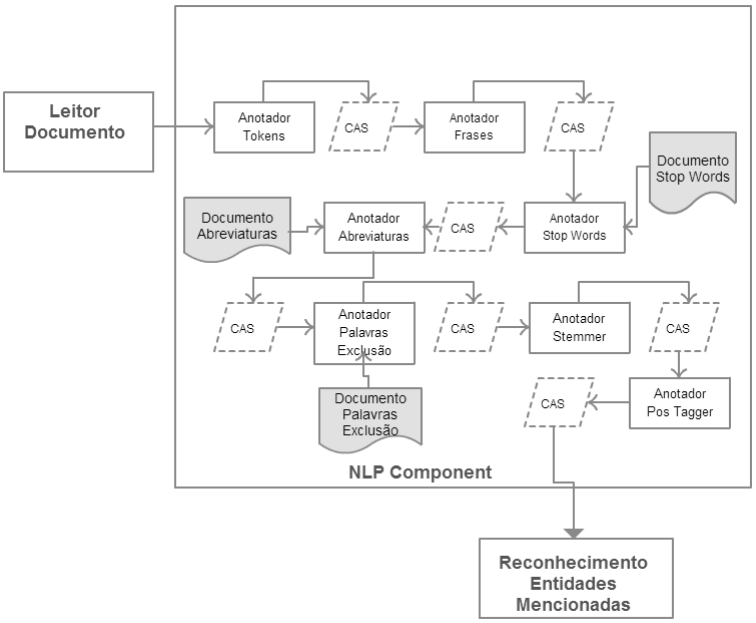

Figura 5.1: Fluxo do Módulo de Princípios de Processamento de Linguagem Natural

palavras sem relevância de processar; expansão de abreviaturas, que visa expandir as abreviaturas encontradas para futura análise; palavras de exclusão, refere-se a palavras que segundo as regras da nomenclatura remetem para a exclusão do termo clínico presente no documento; *stemming*, onde as palavras são reduzidas à sua raiz e pela parte do discurso onde é identificada a categoria gramatical das palavras. O fluxo deste elemento pode ser observado pela Figura 5.1.

A análise levada a cabo nesta fase começa pela identificação dos *tokens* e das frases. Esta identificação, é realizada a partir de heurísticas básicas, que consideram os espaços em branco e os sinais de pontuação. Nesta etapa, foi utilizado um exemplo providenciado pelo UIMA. Também aqui são identificadas as abreviaturas presentes no texto. Após esta análise são verificadas para todos os *tokens* identificados aqueles que correspondem a *stop-words*. Para a realização desta tarefa foi utilizada uma lista com as palavras correspondentes a *stop-words* para a língua portuguesa disponível no Snowball [snob]. São exemplos de *stop-words* as seguintes palavras: o, a , de , quem, e, também,só.

Segue-se a análise da expansão de abreviaturas. Para a realização desta tarefa é utilizada uma lista com 68 abreviaturas disponível pela Infarmed [inf06]. Alguns exemplos de abreviaturas e expansão de termos são os seguintes: HTA – hipertensão arterial e IH – insuficiência hepática.

Sequencialmente são identificadas as palavras de exclusão. Segundo as normas de codificação existem três géneros de palavras que excluem o termo clínico presente no texto, são elas: palavras de negação, incerteza e relativas ao passado. Isto acontece porque a codificação apenas opera sobre as doenças que estão presentes no momento da consulta médica. A lista de palavras de negação utilizada para este trabalho foi: não, nem, sequer, nunca, excepto, suspende. Quanto às palavras de incerteza utilizadas, estas são: suspeito, suspeita, provável, provavelmente, possivelmente, du-

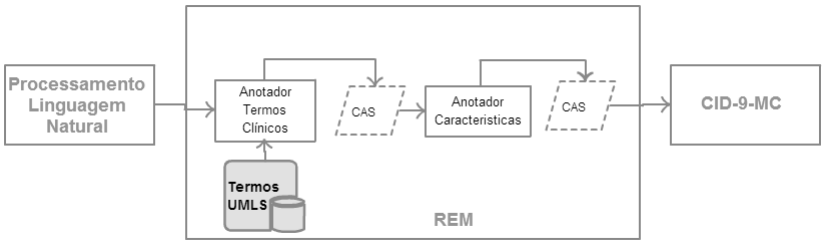

Figura 5.2: Fluxo do Módulo de Reconhecimento de Entidades Mencionadas

vidoso, duvida, talvez. Por fim, a lista de palavras utilizadas que referenciam o passado, são as seguintes: antecedente, antecedentes, passado, antepassado, história, histórico, prévio.

Na fase de *stemming*, na qual se pretende reduzir a palavra à sua raiz, é utilizado o algoritmo de *stemming* disponível pelo Snowball para a língua portuguesa [snoa].

Nesta componente a última fase de processamento passa pelo *part-of-speech* (parte do discurso), onde se pretende conhecer a categoria gramatical das palavras contidas no relatório. Para a realização desta componente é utilizada um complemento do UIMA o anotador do Hidden Markov Model [hmm06]. O modelo de Markov tem como objetivo calcular a sequência de *tags* mais provável na frase. Este cálculo pode ser expresso pela Expressão 5.1 [Bri95], onde: *word*, se refere à palavra à qual se quer conhecer a categoria gramatical; *tag*, corresponde a uma categoria gramatical possível e *previous.n.tag*, corresponde atribuições anteriores. Para a realização desta tarefa, o modelo recebe informação estatística de um ficheiro modelo.

$$maxHMM(word,tag) = Prob(word,tag) * Prob(tag,previous.n.tags) \tag{5.1}$$

5.5 Reconhecimento de Entidades Mencionadas

O objetivo deste trabalho é a atribuição de códigos segundo a nomenclatura CID-9-MC a relatórios clínicos. Para a realização dessa tarefa é essencial perceber quais os termos clínicos que estão presentes no documento.

A tarefa de reconhecimento de entidades mencionadas (REM) é responsável pela identificação das diferentes entidades no texto. Neste trabalho, as diferentes entidades que se visam identificar no relatório são: os termos clínicos relevantes, ou seja, o nome de doenças, sintomas ou diagnósticos e as características associadas a estes termos. Como estes elementos correspondem a entidades diferentes, foram criados diferentes anotadores para o seu processamento. Para a identificação dos termos clínicos é preciso dotar o sistema desse conhecimento, tendo sido usado para tal o recurso fornecido pelo o UMLS.

O fluxo desta componente pode ser observado na Figura 5.2.

5.5.1 Anotador de Termos Clínicos

O objetivo deste anotador é fazer corresponder os termos encontrados no texto com os termos presentes no UMLS de forma a verificar se os termos contidos no texto pertencem ou não a nomes de doenças, sintomas e diagnósticos. Da parte do UMLS, o grupo que contém os termos pretendidos denomina-se Grupo de Doenças e é composto por termos pertencentes a: anormalidades adquiridas; anormalidades anatómicas; bactérias, disfunção molecular e celular; anormalidade congénita; doenças ou síndromes; descobertas; ferimento ou envenenamento; disfunção comportamental ou mental; processo neoplástico; função patológica e sinais ou sintomas.

Para a realização desta tarefa, todos os tokens não anotados até ao momento, ou seja, que não corresponderam a *stop-words* são palavras candidatas para análise. As palavras candidatas foram organizadas em conjuntos. Esta organização foi realizada da seguinte forma, se a palavra se encontrava isolada formava um conjunto simples, porém se a palavra era precedida de mais palavras candidatas, então todas essas palavras construíam um conjunto múltiplo.

De forma a verificar se os termos presentes no texto eram realmente termos clínicos, foram realizadas *queries* ao UMLS. É importante referir que todos os dados, tanto os que derivam do relatório clínico bem como o que resulta do UMLS sofrem um processo de normalização, onde são substituídos os caracteres especiais e substituídos os caracteres em maiúscula para letra minúscula. Também, às expressões vindas do UMLS são removidas as *stop-words*.

As *queries* realizadas ao UMLS começam por verificar o primeiro termo de cada conjunto, se for encontrado o primeiro termo, são procurados os termos seguintes, para o caso dos conjuntos múltiplos. Como se sabe, os termos clínicos podem ter alguma variação, como por exemplo variação de género, é exemplo disso: hipertenso e hipertensa. Devido a esta possível variação é considerada uma distância de edição, não se procurando um termo rigidamente. A distância de edição utilizada foi a distância de Levenshtein [SCA06]. Esta permite que se trabalhe com termos de diferentes comprimentos e permite a realização de três operações: adição, remoção e substituição. O uso da distância de edição é utilizada para calcular o grau de semelhança entre duas ocorrências Expressão 5.2. Na expressão de semelhança: c, corresponde ao termo clínico presente no texto; r, ao termo clínico vindo do UMLS.

$$Semelhanca(c,r) = 100 \times \frac{c.comprimento + r.comprimento}{Distancia - Levensthein(c,r) + c.comprimento + r.comprimento} \tag{5.2}$$

É de notar que para a escolha do termo clínico a associar ao termo encontrado no relatório têm em conta dois elementos, o grau de semelhança e o número de *tokens* que compõe o conjunto. A nível do grau de semelhança nunca pode ser inferior a 85%, e o conjunto com mais palavras é preferível.

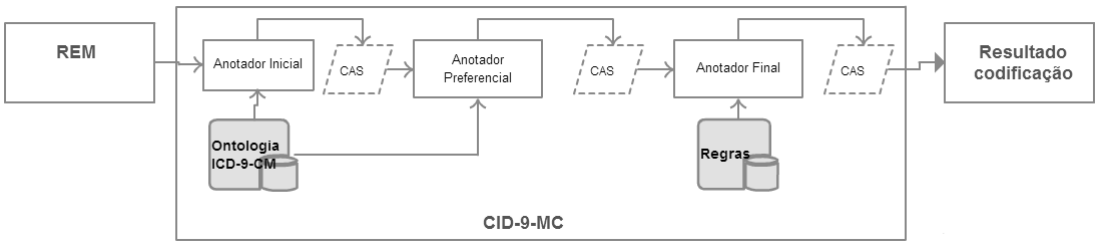

Figura 5.3: Fluxo do Módulo CID-9-MC

5.5.2 Anotador de Características

Este anotador pretende identificar todas as características associadas ao termo clínico presente do documento. Para esta tarefa as anotações realizadas previamente (*stop-word* e termos clínicos) não são considerados. As palavras candidatas correspondem a palavras que surgem após um termo clínico e pertencem à categoria gramatical de advérbio ou adjectivo. Neste caso, todas as palavras na situação descrita são anotadas como características.

5.6 CID-9-MC

O objetivo desta componente é a realização da atribuição de códigos na nomenclatura CID-9-MC ao relatório clínico. Como já foi referido esta é uma tarefa complexa e não trivial. Para a atribuição de um código a um relatório várias são as etapas que se têm que realizar. Neste trabalho, foram consideradas três etapas subsequênciais, são elas:

Inicial: o objetivo deste anotador é a descoberta do código da categoria CID-9-MC associado ao termo clínico.

Preferencial: este anotador visa perceber se o estado do termo clínico no texto, se está realmente presente ou não, organizando por isso os termos em três listas: excluídos, omitidos e presentes. É também este o anotador responsável pela pesquisa do melhor código a atribuir, ou seja, responsável pela verificação das subcategorias e subclassificações do código e pelo cumprimento das regras gerais de codificação.

Final: este anotador recebe o resultado provido do anotador anterior e sobre os dados recebidos é responsável pelo cumprimento das regras específicas da codificação.

O fluxo desta componente pode ser observado na Figura 5.3.

5.6.1 Anotador Inicial

Analogamente às regras de codificação, este anotador pretende "Encontrar o termo no índice alfabético". Este anotador tem como objetivo encontrar o código da categoria CID-9-MC associado ao termo clínico encontrado no relatório. Para elaborar a sua tarefa, este anotador recebe os

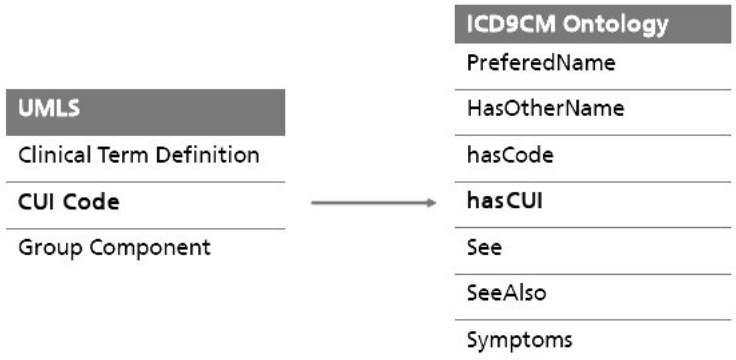

Figura 5.4: Elemento em comum entre o UMLS e a Ontologia CID-9-MC

dados provindos do anotador de termos clínicos a fim de saber quais as ocorrências a analisar; e utiliza como fonte de conhecimento a ontologia CID-9-MC.

A descoberta do código da categoria CID-9-MC associado ao termo clínico presente no relatório é uma tarefa trivial, isto porque, por parte do UMLS e da Ontologia existe um elemento em comum que corresponde ao código identificador da doença (CUI, como pode ser observado pela Figura 5.4).

5.6.2 Anotador Preferencial

De acordo com as etapas realizadas na codificação manual e de acordo com as regras de codificação, este anotador compreende as seguintes etapas: "Pesquisa de modificadores ao termo principal", onde neste caso o termo principal é o termo clínico e os modificadores correspondem às negações, incertezas e palavras que remetem para o passado; e a "Pesquisa de notas associadas ao termo principal", que neste caso em particular se refere às características associadas ao termo clínico.

O fluxo deste anotador pode ser visualizado pela Figura 5.5. O objectivo deste anotador é o cumprir com as regras gerais intrínsecas ao processo de codificação. Para a realização desta etapa, o resultado dos anotadores palavras de exclusão, características e o anotador inicial são requeridos.

Primeiramente, este anotador começa por verificar se os termos clínicos encontrados no documento correspondem a sinais ou sintomas. Isto é realizado através do código associado ao termo clínico, pois, segundo as regras de codificação os códigos dos sinais e sintomas encontram-se entre os códigos 780.0 ao 799.9. Esta verificação faz parte de uma das regras gerais, que nos indica que se num documento estiver presente uma doença e um sintoma a ela associado, então o código do sintoma tem que ser excluído.

A fase seguinte visa verificar o estado dos termos clínicos no texto, isto porque, por exemplo na frase "O paciente não apresenta pneumonia", a doença pneumonia não pode ser codificada pois não faz parte do diagnóstico. Nesta fase existem duas formas de excluir o termo clínico do texto, uma deriva das palavras de negação e incerteza e a outra das palavras que se referem ao passado.

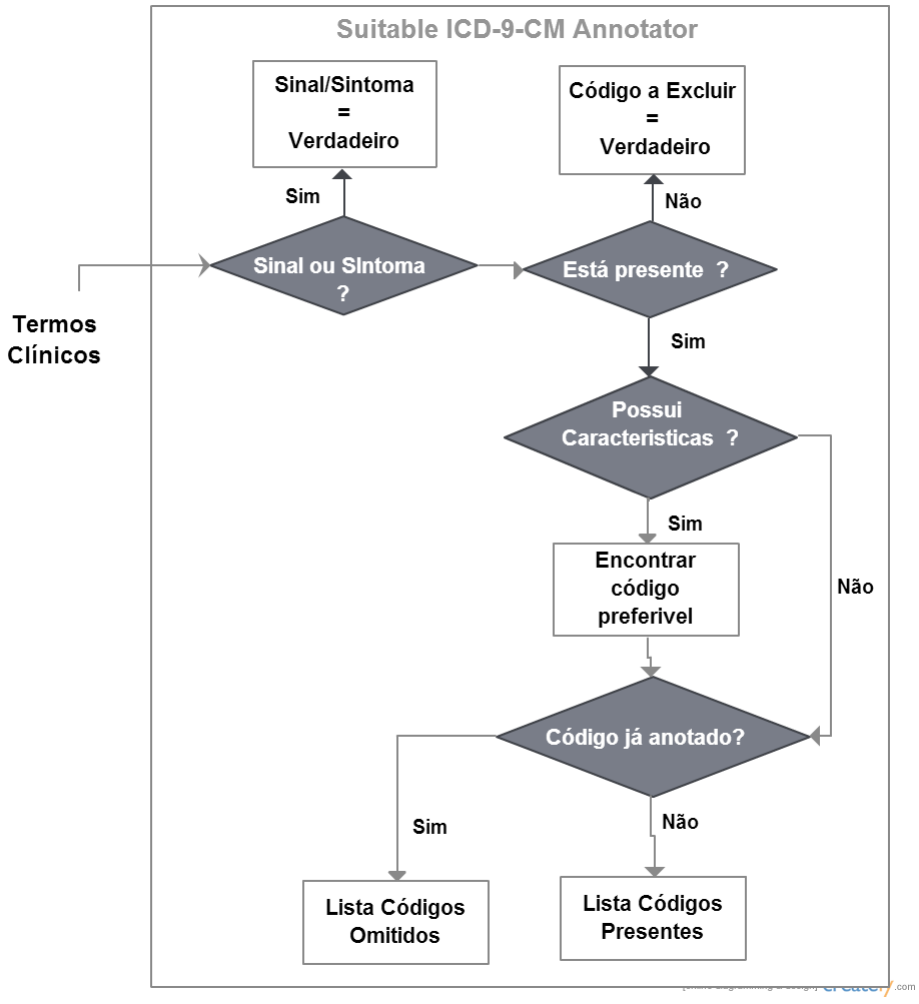

Figura 5.5: Fluxo da componente preferencial

Na primeira situação, pela observação dos relatórios clínicos elaborados como conjunto de teste, verificou-se que o termo clínico aparece no máximo numa distância de cinco palavras entre a palavra de negação e incerteza e o termo clínico. Em relação ao segundo caso, pela análise dos relatórios verificou-se que quando aparece uma palavra que remete para o passado surge uma lista de termos clínicos evidenciados anteriormente, considerando-se neste caso, não uma distância de cinco palavras, mas a exclusão de todos os termos clínicos na frase que apareceram após a ocorrência de uma palavra que remete para o passado. Todos os termos excluídos por este processo vão para a lista de termos excluídos.

Depois desta análise, para todos os códigos detetados são verificadas as suas características, isto porque, as características ajudam na definição do melhor código a atribuir. A análise do melhor código a atribuir é realizada através da técnica de alinhamento de paráfrases. Para isso, é recebido da ontologia todas as definições dos termos associados aos códigos das subcategorias e subclassificações. Esses termos recebidos são normalizados, ou seja, os caracteres especiais são substituídos, as letras são convertidas para minúsculas e as palavras correspondentes a *stop words* são removidas, pois o seu significado não acrescenta valor à definição do termo clínico. Depois de feita a normalização, é verificado qual a definição que contém mais elementos contidos nas características, de forma a determinar o melhor código. Esta fase corresponde ao "Nível de detalhe" presente nas regras gerais, esta indica-nos que o termo associado a um termo clínico deve ser sempre com o maior número de dígitos possíveis.

Por fim, os resultados são anotados com as suas características (See, SeeAlso e Lista de Sintomas). No caso de já se ter verificado a ocorrência de um determinado código este vai para a lista de códigos omitidos, porque o código não necessita de ser processado duas vezes. No caso de ser a primeira ocorrência do código, então este vai para a lista de presentes.

5.6.3 Anotador Final

O objetivo deste anotador é cobrir todas as regras específicas da codificação. Porém, neste anotador ainda são verificadas algumas regras gerais, na medida que só neste estado é que se sabe exatamente os códigos que o relatório contém e o seu estado. Para o desenvolvimento deste anotador é necessário o resultado do anotador anterior a ontologia CID-9-MC e do sistema de regras desenvolvido.

Primeiramente, este ponto verifica os seguintes aspetos presentes nas regras de codificação: "Condições que fazem parte integrante da doença" e "Condições que não fazem parte integrante da doença". Para tal, o sistema verifica se existem sintomas presentes no documento clínico, se existirem verifica em todas as doenças presentes se estas estão relacionadas com o sintoma. No caso da doença estar relacionada com o sintoma, então o código do sintoma é excluído, se não o código mantém-se.

Após a verificação efectuada, o sistema verifica as regras específicas de codificação. Para isso só necessita de enviar para a componente de regras as listas de códigos omitidos e presentes.

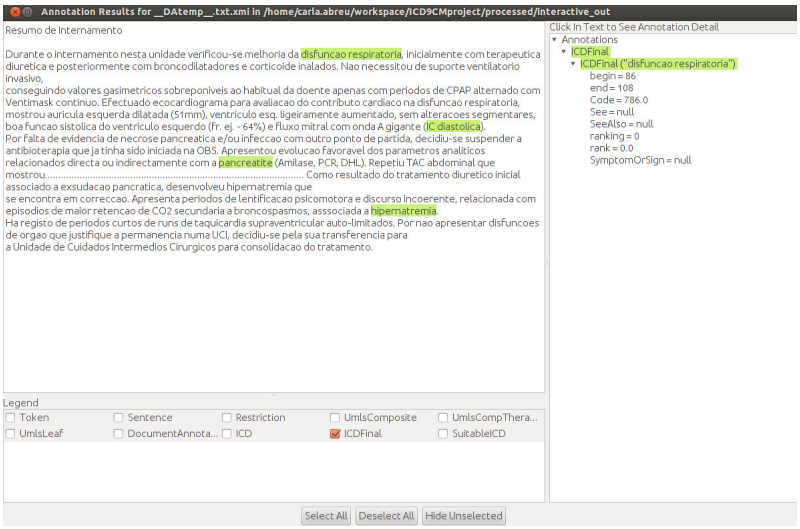

Figura 5.6: Visualização do resultado pela visualização providenciada pela UIMA

5.7 Atribuição de Códigos

Esta componente é responsável pela extração dos resultados provenientes das análises elaboradas anteriormente e entregar essa informação ao utilizador. Neste sistema existem duas formas de entregar o resultado da codificação ao utilizador: utilizando a visualização providenciada pelo UIMA (Figura 5.6) e por um grafo (Figura 5.7).

A visualização providenciada pelo UIMA permite a visualização integral do relatório clínico processado e as anotações realizadas. Neste género de visualização, o utilizador pode selecionar a tipo de anotação que quer ver no documento, ao selecionar um tipo de anotações, os elementos apareceram sublinhados com uma cor. Caso o utilizador queira perceber as características da anotação, terá apenas que carregar sobre um elemento e do lado direito da janela de visualização aparecerão as características do mesmo.

De forma a facilitar a compreensão do codificador de todas as etapas realizadas pelo sistema

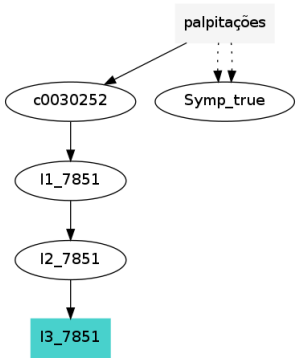

Figura 5.7: Gráfico com as três iterações da codificação

até chegar ao código final foi elaborada uma representação gráfica. Nesta representação os termos clínicos aparecem representados com rectângulo sombreado de cinzento; a azul aparece o código CID-9-MC sugerido; a cinzento escuro é representado a definição da regra especifica evidenciada; a rosa aparece um termo que caracteriza um sinal ou sintoma; a vermelho são representados os códigos das características See, SeeAlso.

Capítulo 6

Avaliação

Este capítulo visa apresentar a avaliação realizada ao sistema desenvolvido. Este sistema teve duas formas de avaliação: uma objectiva e outra subjectiva. A primeira avaliação foi elaborada de uma forma quantitativa, para tal foram utilizadas dois portais: o portal de terminologias e ontologias de saúde (HeTop) e o MediLexion. A segunda forma de avaliação correspondeu a uma opinião de um especialista em codificação.

Neste capítulo serão descritos os seguintes elementos: o conjunto de dados utilizados; as métricas de avaliação; e o Humano na avaliação. Por fim, são apresentados os resultados obtidos e uma análise aos mesmos.

6.1 Conjunto de Dados

Como foi referido em capítulos anteriores os relatórios clínicos contém informação confidencial, pelo que, por aspectos éticos e sociais os mesmos não se encontram acessíveis.

Este sistema contou com o apoio de dois médicos portugueses que criaram os documentos para avaliação. Os relatórios foram desenvolvidos para a área de cardiologia em forma de texto livre. Estes documentos foram realizados em diferentes formatos, e continham uma vasta nomenclatura médica contendo também algumas abreviaturas.

Para a avaliação foram criados no total 22 relatórios clínicos, com uma média de 336 palavras por documento.

6.2 Métricas de Avaliação

Este sistema foi avaliado em três situações:

1. Verificação dos termos extraídos pela sistema;

2. Verificação dos identificadores únicos associados aos termos clínicos;

3. Verificação da codificação realizada.

A métrica utilizada para avaliar esta situação foi a precisão (P), cuja equação pode ser observada em X.

$$P = \frac{verdadeirosPositivos}{verdadeirosPositivos + falsosPositivos} \tag{6.1}$$

A situação 1 pretende verificar se os termos extraídos pelo sistema são realmente termos clínicos. Para esta situação, os verdadeiros positivos correspondem aos termos extraídos pelo sistema como sendo termos clínicos e que realmente o são; e os falsos positivos são os termos extraídos pelo sistema como sendo termos clínicos mas que na realidade não o são.

O objetivo da situação 2 é verificar se os identificadores únicos dos termos clínicos estão a ser atribuídos de forma correcta. Nesta situação, os verdadeiros positivos correspondem a uma correta correspondência entre o termo clínico e o seu identificador único realizado pelo sistema, enquanto os falsos positivos correspondem a atribuições incorretas.

Por fim, a situação 3 visa verificar se o código CID-9-MC atribuído ao termo clínico faz sentido. Aqui, os verdadeiros positivos correspondem aos casos em que o termo clínico está bem associado ao código atribuído pelo sistema e os falsos positivos correspondem aos casos em que o código atribuído pelo sistema ao termo clínico não faz sentido.

A avaliação dos três casos foi feita por inspeção manual, através do uso para a situação 1 e 2 do HepTop e para a situação 3 do MediLexion.

É de notar que só foi possível a utilização desta métrica na avaliação por falta de conhecimento da terminologia clínica.

6.3 O Humano no Processo

O sistema desenvolvido neste trabalho consiste num sistema de apoio à decisão. Como foi referido nos capítulos anteriores, o resultado da codificação é sempre determinado por um codificador.

De forma a avaliar o sistema, foi pedido a um especialista em codificação que avaliasse o sistema. Para isso foram compilados os resultados obtidos pelo sistema para 16 relatórios clínicos. Este documento teve também a descrição de todos os elementos contidos no mesmo.

6.4 Resultados e Análise

A Tabela 6.1 representa os resultados obtidos da avaliação desenvolvida. Esta avaliação indica, para o caso 1, que apenas 3% de 94 termos extraídos pelo sistema não correspondiam a termos clínicos. Quanto à situação 2, foi verificado que 9,5% dos identificadores únicos atribuídos aos termos clínicos estavam incorretos.

Quanto ao processo de codificação, a situação 3, o sistema indicou que 86,17% dos códigos atribuídos aos termos clínicos tinham relação.

Situação	verdadeirosPositivos	falsosPositivos	Total	Precisão
1	91	3	94	96.81%
2	85	9	94	90.43%
3	81	13	94	86.17%

Tabela 6.1: Resultados da avaliação ao sistema

Infelizmente não foi possível comparar os resultados obtidos por este sistema com os sistemas desenvolvidos na área e pois: as métricas de avaliação utilizadas foram distintas e as situações foram também analisadas de forma distinta.

Como foi referido, o resultado da codificação é sempre atribuído por um especialista em codificação. É também este profissional que determina o desempenho de sistemas na área. Para se perceber se o sistema desenvolvido seria aceite, foi enviado um conjunto de relatórios clínicos anotados pelo sistema a um codificador, porém, só foi possível obter o resultado para relatório clínico apresentado na Tabela 6.2, cujo resultado se pode observar na Tabela 6.3. O resultado obtido é muito vago para que se possa obter uma opinião concreta sobre o desempenho do sistema.

A Sra.D. “Joaquina Alves” é uma doente de 60 anos de idade, autónoma, doméstica. Apresenta como fatores de risco cardiovascular hipertensão arterial e sobrecarga ponderal. Tem antecedentes de perturbação depressiva, hipotireoidismo iatrogénico secundário a tireoidectomia total em contexto de bócio multinodular. Não apresenta outros antecedentes patológicos relevantes. Nega alergias conhecidas. Encontra-se medicada em ambulatório com Fluoxetina 20 mg qd, Perindopril 8 mg qd, Levotiroxina 0.1 mg qd. Permaneceu assintomática do foro cardiovascular até ao dia 24 de Abril de 2013 quando, após durante celebração fúnebre do seu filho, iniciou quadro de dor retroesternal opressiva, associada a dispneia seguida de síncope com recuperação espontânea em poucos minutos. Foi transportada ao Serviço de Urgência deste hospital onde deu entrada assintomática. O ECG inicial mostrava ritmo sinusal, 50 bpm, má progressão R e inversão profunda da onda T nas derivações precordiais. O ecocardiograma revelou depressão moderada da função ventricular esquerda, com acinésia do ápex e segmentos apicais e hipercinésia dos segmentos basais. Analiticamente foi documentada elevação dos marcadores de necrose miocárdica (troponina I de 3.0 ng/mL), sem outras alterações relevantes. A doente ficou internada no Serviço de Cardiologia, onde apresentou evolução clínica favorável, sem recorrência sintomática e mantendo estabilidade elétrica e hemodinâmica. O cateterismo cardíaco revelou ausência de doença coronária angiográfica significativa. Ao 5.º dia de internamento foi repetido ecocardiograma que mostrou recuperação da função sistólica com regressão das alterações da motilidade segmentar. Foi assumido o diagnóstico de miocardiopatia de stress e a doente teve alta, orientada para o Médico de Família e para a Consulta de Cardiologia, com indicação para realização posterior de ressonância magnética cardíaca. Medicação proposta: AAS 100 mg qd, Rosuvastatina 10 mg qd, Perindopril 5 mg qd, Fluoxetina 20 mg qd, Levotiroxina 0.1 mg qd.

Tabela 6.2: Relatório Clínico Avaliado por o codificador

Descrição Termo Clínico	Código atribuido Codificador	Código Atribuido pelo sistema
Miocardiopatia de stress	425.9	425
Hipertensão Arterial	401.9	997.9
Sobrecarga ponderal	278.02	-
Hipotiroidismo pós-cirúrgico	244.0	-
Dispneia	786.09	786.09
Síncope	780.02	780.02

Tabela 6.3: Resultado do processo de codificação 6.2

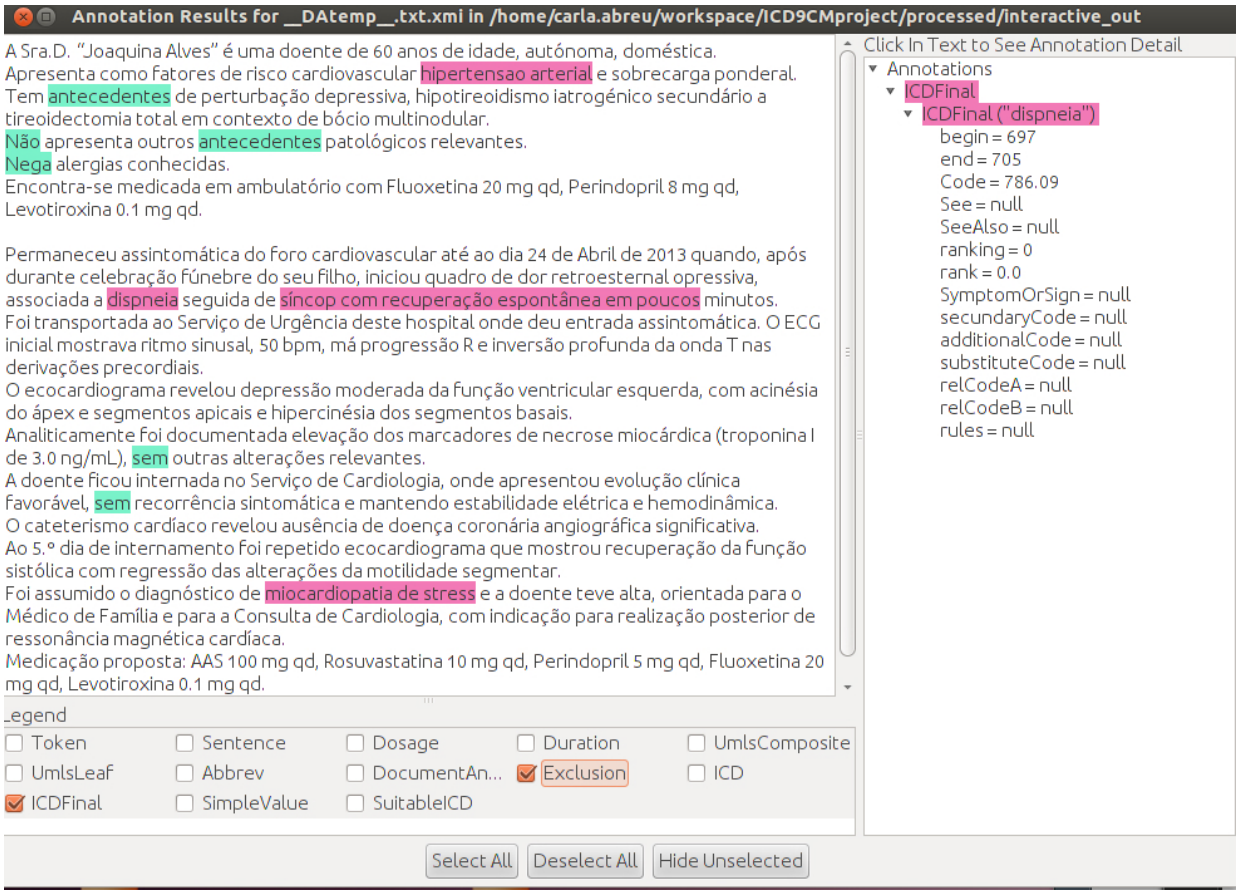

Figura 6.1: Resultado do relatório clínico avaliado

Capítulo 7

Conclusões

Diariamente são produzidos vários volumes de relatórios clínicos nos Hospitais. Grande parte destes documentos são já produzidos no formato eletrónico.

Os dados contidos nos relatórios clínicos são muito úteis, pois servem como base a estudos epidemiológicos e permitem que os Hospitais recebam verba externa [CL95], imperativa ao seu funcionamento. Porém, a estrutura destes documentos não segue nenhum padrão e ao mesmo tempo são escritos de forma livre. Estes dois elementos combinados fazem com que a tarefa de processamento dos documentos seja uma tarefa bastante complexa [CW07].

De forma a ser percetível o conteúdo do documento, entende-se por processamento a forma de extrair do mesmo os dados significantes, como os termos relativos a doenças, sintomas e diagnósticos e a indexação aos mesmos de um ou mais códigos que remetam ao conteúdo nele contido. Neste trabalho, foi utilizado a nomenclatura CID-9-MC [MR09]. Este sistema possui centenas de códigos e um vasto conjunto de regras associados à tarefa de codificação, tornando esta tarefa morosa e complexa [CW07].

Esta tese apresenta um sistema de codificação que visa: numa primeira fase, extrair termos clínicos como nomes de doenças, sintomas e diagnósticos do texto; e numa segunda fase, codificar os termos encontrados de acordo com as regras vigentes.

O sistema de codificação foi projetado em seis componentes: o pré-processamento do documento; a leitura do documento; o processamento de linguagem natural; o reconhecimento de entidades mencionadas; a codificação em CID-9-CM e por fim o processamento do resultado para o entregar ao utilizador. Este sistema foi desenvolvido sobre uma arquitetura de gestão de informação desestruturada (UIMA).

Para a avaliação deste sistema, foram criadas por dois médicos especialistas em cardiologia 22 cartas de alta. Os resultados obtidos apontam para uma precisão de 96.81% na captura de termos clínicos; uma precisão de 90.43% na atribuição do identificador único ao termo clínico. Quanto à codificação final atribuída aos termos clínicos, em 86.17% das vezes o código estava relacionado com a doença.

7.1 Trabalho Futuro

Para que este sistema esteja adaptado a todos os capítulos da nomenclatura de classificação (representados na Tabela 3.1) poderão ser programadas as regras específicas para os capítulos em falta. A programação destas passará pelo esforço na compreenção dessas regras, visto que a nível da introdução das mesmas no sistema se faz de uma forma simples devido ao uso da framework Java Expert Drools Engine.

A nível do tratamento do texto, uma melhoria a fazer ao sistema desenvolvido passaria pelo tratamento semântico do texto. Isto permitia o tratamento das palavras de exclusão de uma forma mais ponderada.

Como é evidente, a nível da avaliação este trabalho necessitaria de uma avaliação mais extensiva, para a qual se deveriam utilizar as métricas de abrangência e a medida F. Esta avaliação, só podetá ser realizada com o apoio de um codificador, devido à familiaridade com que o mesmo tem com o tema em questão.

Referências

[AFR97] R. Fikes A. Farquhar e J. Rice. The ontolingua server: a tool for collaborative ontology construction. international journal of human-computer studies, 1997.

[Agu12] Bruno Aguiar. *Information Extraction From Medication Leaflets*. PhD thesis, University of Oporto, 35-48, Portugal, 2012.

[AM06] Sophia Ananiadou e John McNaught. Text mining for biology and biomedicine. 2006.

[Bri95] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging, 1995.

[BSR97] K. Knight B. Swartout, P. Ramesh e T. Russ. Toward distributed use of large-scale ontologies. in aaai symposium on ontological engineering, 1997.

[CL95] R. Baud J.R. Scherrer C. Lovis, P.A. Michel. Use of conceptual semi-automatic icd-9 encoding system in an hospital environment. 1995.

[CW07] Abe Coffman e Nat Wharton. Clinical natural language processing auto-assigning icd-9 codes. 2007.

[Dom98] J. Domingue. Tadzebao and webonto: Discussing, browsing, editing ontologies on the web. in 11th knowledge acquisition for knowledge-based systems workshop, 1998.

[Dro13] Drools - the business logic integration platform. `http://www.jboss.org/drools/`, Last checked: 05/06/2013, 2013.

[Fer11] Liliana Ferreira. *Medical Information Extraction in European Portuguese*. PhD thesis, University of Aveiro, 62-65 81-101, Portugal, 2011.

[FL04] D. Ferrucci e A. Lally. Uima an architectural approach to unstructured information processing in the corporate research environment. natural language engineering, 2004.

[FS08] Richárd Farkas e Gyorgy Szarvas. Automatic construction of rules-based icd-9-cm coding system. 2008.

[Gra13] Graphviz - graph visualization software. `http://www.graphviz.org/`, Last checked: 05/06/2013, 2013.

[hmm06] Hidden markov model tagger annotator. `http://uima.apache.org/sandbox.html#tagger.annotator`, Last checked: 05/06/2013, 2006.

REFERÊNCIAS

[ICD10] Icd-9-cm official guidelines for coding and reporting. http://www.cdc.gov/nchs/data/icd9/icdguide10.pdf, Last checked: 05/06/2013, 2010.

[IGU07] A. Arzumtsuan I. Goldstein e Ozlem Uzuner. Three approaches to automatic assignment of icd-9-cm codes to radiology reports. 2007.

[inf06] Prontuário terapêutico. http://m.infarmed.pt/Prontuario/Normas.aspx, Last checked: 05/06/2013, 2006.

[Jen13] Apache jena. http://jena.apache.org/, Last checked: 05/06/2013, 2013.

[JGM05] Valerie Watzlaf Jennifer Garvin e Sohrab Moeini. Development and use of automated coding software of enhance antifraud activities. 2005.

[JHGT02] R. W. Fergerson W. E. Grosso M. CrubOzy H. Eriksson N. F. Noy J. H. Gennari, M. A. Musen e S. W. Tu. The evolution of protégé: An environment for knowledge-based systems development. international journal of human-computer studies, 2002.

[JMB11] Jesualdo Breis Fancisco Sanchez Juana Martinez, Rafael Garcia e Rodrigo Bejar. Ontology learning from biomedical natural language documents using umls. 2011.

[JP08] Pawel Matykiewicz DJ Hovermale Neil Johnson Bretonnel Cohen Wlodzislaw John Pestian, Christopher Brew. A shared task involving multi-label classification of clinical free text. 2008.

[JPW06] Yitao Zhang Jon Patrick e Yefeng Wang. Developing feature types for classifying clinical notes. 2006.

[KFT98] A. Tamura K. Fukuda, T. Tsunoda e T. Takagi. Toward information extraction: Identifying protein names from biological papers. 1998.

[KP07] Anne Kao e Stephen Poteet. *Natural Language Processing and Text Mining, 1-3*. 2007.

[KR08] Manuela Kunze e Dietmar Rösner. Uima for nlp based researchers' workplaces in medical domains. in proceedings ofworkshop 'uima for nlp', 2008.

[Kum11] Ela Kumar. *Natural Language Processing*. 2011.

[KVH09] C. Roeder K. Verspoor, W. Baumgartner e L. Hunter. Abstracting the types away from a uima type system, 2009.

[MR97] R. Besançon M. Rajman. Text mining: Natural language techniques and text mining applications. 1997.

[MR09] Enrique Mota e Arturo Romero. Auto coder, web coder: Sistemas de informatión para la codificatión sanitaria. 2009.

[Mus99] Ion Muslea. Extraction patterns for information extraction tasks: A survey. 1999.

[oM12] U.S. National Library of Medicine. Fact sheet: The national library of medicine. http://www.nlm.nih.gov/pubs/factsheets/nlm.html, Last checked: 04/06/2013, 2012.

[Pro13] Protégé. http://protege.stanford.edu/, Last checked: 05/06/2013, 2013.

REFERÊNCIAS

[RF07] James Sanger Ronen Felman. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, 1-13 40*. 2007.

[RJ99] Ellen Riloff e Rosie Jones. Learning dictionaries for information extraction by multi-level bootstrapping. 1999.

[SCA06] A. Freeman S. Condon e C. Ackerman. Cross linguistic name matching in english and arabic: A 'one to many mapping' extension of the levenshtein edit distance algorithm, 2006.

[snoa] Portuguese stemming algorithm. `http://snowball.tartarus.org/algorithms/portuguese/stemmer.html`, Last checked: 05/06/2013.

[snob] A portuguese stop word list. `http://snowball.tartarus.org/algorithms/portuguese/stop.txt`, Last checked: 05/06/2013.

[SWD05] Tong Zhang Sholom Weiss, Nitin Indurkhya e Fred Damerau. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. 2005.

[TPSJNH06] MD Tammy Powell Stuart J. Nelson e Betsy L. Humphreys. The unified medical language system (umls) project. `http://www.nlm.nih.gov/archive/20130426/mesh/umlsforelis.html`, Last checked: 04/06/2013, 2006.

[TR03] Marcelo Fisman Thomas Rindflesch. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. 2003.

[Zia] Ramon Ziai. A flexible annotation-based architecture for intelligent language tutoring systems. `http://www.sfs.uni-tuebingen.de/~rziai/papers/Ziai-09.pdf`, Last checked: 24/04/2013.