

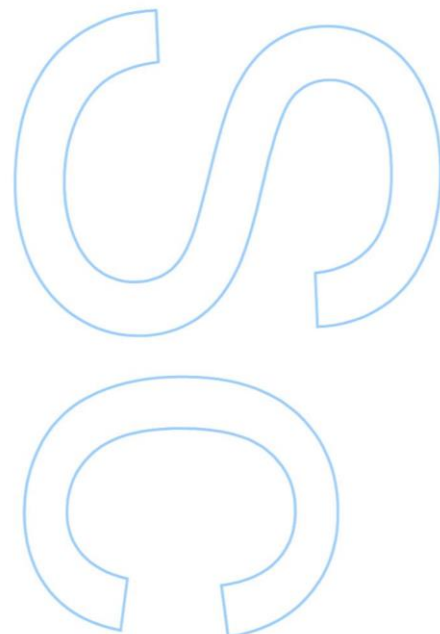
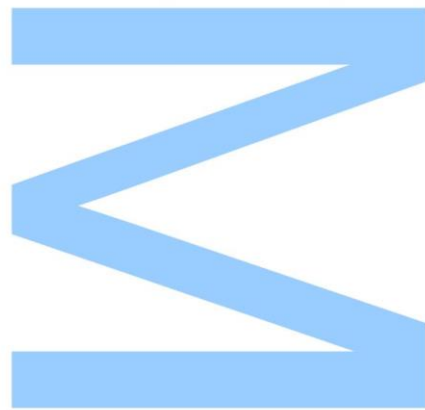
Identificação de termos relevantes em relatórios usando *text mining*

Pedro da Silva Bastos

Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos
Departamento de Ciência de Computadores
2017

Orientador

Alípio Mário Guedes Jorge, Professor Associado,
Faculdade de Ciências da Universidade do Porto

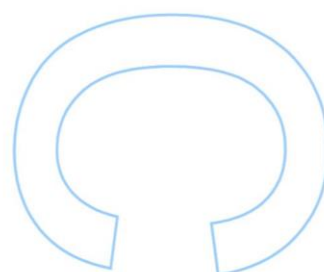
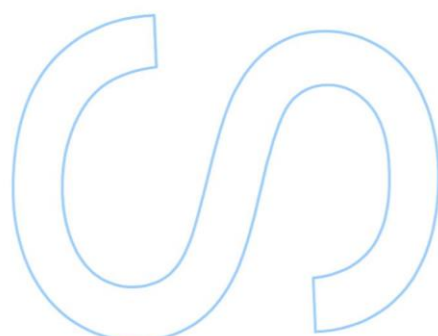
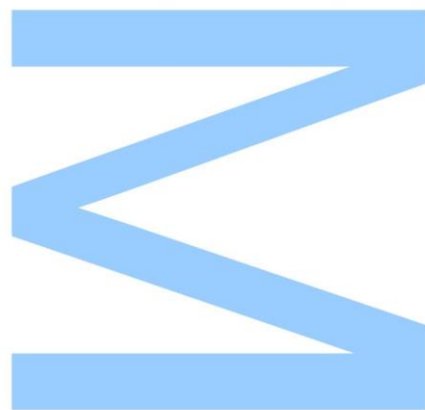




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/____



Aos meus pais

Agradecimentos

Neste espaço gostaria de agradecer às pessoas que me acompanharam ao longo do desenvolvimento da dissertação, bem como, e não poderia esquecer, todos aqueles que me acompanharam durante o meu processo de formação.

Ao Prof. Alípio Mário Guedes Jorge, na qualidade de orientador, por todo o apoio dado durante a realização da dissertação.

À *FlipKick* e ao Nuno Tavares, na qualidade de CEO da *FlipKick*, por disponibilizar todo o material indispensável à realização da tese.

À Mafalda Tavares e Prof^a. Fátima Oliveira pelas diversas revisões ortográficas.

Aos meus professores e amigos que me acompanharam ao longo do meu processo de formação.

Ao meu país, aos portugueses, às portuguesas e a todos aqueles que, direta ou indiretamente, lutaram por um ensino superior público.

À minha família.

Aos meus pais, a quem estarei eternamente grato, pelo esforço, educação e crença que depositaram em mim.

“Ter que pagar pelos próprios sonhos deve ser o pior dos desesperos” — José Saramago, A Viagem do Elefante

Abstract

A current challenge in the area of medical information is the need to treat patients' clinical data more efficiently. Much of the information that appears in a clinical report is in an unstructured format, and normally it is written in the narrative in form of natural language. All these natural language barriers make the medical information management process difficult. However, automatic tasks can make the process easier. An example is the automatic extraction of relevant terms, in clinical reports, such as terms referring to pathologies.

In this work, we intended to create a system that is capable of automatically extracting clinical conditions, descriptions of clinical conditions and areas of incidence of clinical conditions in thyroid reports. The system aims, in the first phase, to identify the relevant occurrences, in the second phase, the identification of the entities and, in the third phase, the automatic extraction of relevant terms based on the entities association with the relevant occurrence's. It is comprised of eight different components; pre-processing of clinical text, identification of possible relevant occurrences, identification of relevant statements, part-of-speech-tagging and lemmatization of relevant statements, named entity recognition, sequences of the entities identified, the association of the entities to the relevant occurrences and the extraction of relevant terms.

A total of 1690 thyroid reports was available for review. 200 randomly chosen reports were annotated in order to evaluate the effectiveness of the system, comparing the results of the automatic extraction with the annotations of the clinical reports. The results obtained indicate an accuracy of 98.9% for the extraction of clinical conditions, a 98.5% accuracy for the extraction of incidence zones of the clinical conditions and a 98.2% F1 for the extraction of descriptions of clinical conditions. The system indicates that 95.4% of the relevant terms are correctly extracted.

Clinical reports are provided by the company *Flipkick*.

Key-words: Information Extraction, Thyroid, Text Mining, Natural Language Processing, Named Entity Recognition, Clinical Reports

Resumo

Um desafio atual que se verifica na área da informação médica é a necessidade de se tratar de forma mais eficiente e com menos recursos os dados clínicos de pacientes. Muita dessa informação que consta nos relatórios clínicos está pouco ou nada estruturada, pois na maior parte das vezes é apresentada na narrativa sob a forma de linguagem natural. Todos estes entraves da linguagem natural tornam difícil o processo de gestão de informação médica. Todavia, a automatização de tarefas pode facilitar o processo. Um exemplo é o da extração automática de termos relevantes referentes a patologias nomeadas em relatórios clínicos.

Neste trabalho, apresentamos um sistema capaz de extrair automaticamente condições clínicas, descrições de condições clínicas e zonas de incidência das condições clínicas em relatórios da tiroide. O sistema visa, numa primeira fase, a identificação de ocorrências relevantes, numa segunda fase, a identificação das entidades e, por fim, a extração automática de termos relevantes com base nas entidades associadas às ocorrências relevantes. O trabalho foi projetado em oito componentes diferentes, sendo elas: o pré-processamento de texto clínico, a identificação de possíveis ocorrências relevantes, a identificação de frases relevantes, o *part-of-speech-tagging* e lematização de frases relevantes, a identificação de entidades, a construção de sequências das entidades identificadas, a associação das entidades às ocorrências relevantes e a extração de termos relevantes.

Estiveram disponíveis, para análise, um total de 1690 relatórios da tiroide. Desses, foram escolhidos aleatoriamente 200 relatórios, que foram anotados de forma a ser possível avaliar a eficácia do sistema, comparando os resultados da extração automática com as anotações dos relatórios clínicos. Os resultados obtidos apontam para uma *accuracy* de 98.9% na extração de condições clínicas, 98.5% para a extração das zonas de incidência das condições clínicas e um F1 de 97.8% para a extração de descrições de condições clínicas. O sistema indica que 95.4% dos termos relevantes são corretamente extraídos.

Os relatórios clínicos foram fornecidos pela empresa *Flipkick*.

Palavras-chave: Extração de informação, Tiroide, *Text Mining*, Processamento de Linguagem Natural, Reconhecimento de Entidades, Relatórios clínicos

Conteúdo

Agradecimentos	iii
Abstract	v
Resumo	vii
Conteúdo	xi
Lista de Tabelas	xiv
Lista de Figuras	xvi
Lista de Blocos de Código	xvii
Acrónimos	xix
1 Introdução	1
1.1 Contexto	1
1.2 Motivação	2
1.3 Objetivo	3
1.4 Contribuições	4
1.5 Estrutura da dissertação	4
2 Ferramentas e conceitos utilizados	5
2.1 <i>Text Mining</i>	5
2.2 Processamento de Linguagem Natural	5

2.3	Extração de informação	6
2.4	Reconhecimento de entidades	6
2.5	Ocorrências relevantes	7
2.6	<i>TreeTagger</i>	7
2.7	<i>Package openNLP</i>	8
3	Estado da Arte	9
3.1	Extração de termos relevantes	9
3.1.1	Processamento de linguagem médica	9
3.1.2	Reconhecimento de entidades	10
3.1.3	Sistemas de extração de termos relevantes	11
4	Solução proposta	13
4.1	Pré-processamento de texto clínico	15
4.2	Identificação de possíveis ocorrências relevantes	17
4.3	Identificação de frases relevantes	18
4.4	<i>POSTagging</i> e lematização de frases relevantes	19
4.5	Reconhecimento de entidades	20
4.5.1	Complemento Circunstancial de Lugar	20
4.5.2	Condições Clínicas	23
4.5.3	Descrições de condições clínicas	24
4.5.4	Exceções	26
4.6	Sequências	27
4.6.1	Ocorrências relevantes	29
4.6.2	Fim de frase	29
4.6.3	Verbos	29
4.6.4	Ordem de ocorrência	30
4.6.5	Obter sequência	32
4.7	Extração automática das entidades associadas às ocorrências relevantes	34

4.7.1	Extração de condições clínicas	35
4.7.2	Extração de descrições de condições clínicas	36
4.7.3	Extração de complementos circunstanciais de lugar	39
4.8	Extração de termos relevantes	41
5	Avaliação	45
5.1	Conjunto de dados	45
5.2	Medidas de avaliação	48
5.3	Resultados e Análise	49
5.3.1	Primeiro conjunto de dados	49
5.3.1.1	Extração de condições clínicas	49
5.3.1.2	Extração de complementos circunstanciais de lugar	50
5.3.1.3	Extração de descrições de condições clínicas	51
5.3.1.4	Extração de termos relevantes	51
5.3.2	Segundo conjunto de dados	52
5.3.2.1	Extração de condições clínicas	52
5.3.2.2	Extração de complementos circunstanciais de lugar	52
5.3.2.3	Extração de descrições de condições clínicas	53
5.3.2.4	Extração de termos relevantes	53
5.3.3	Avaliação do sistema	54
6	Conclusões	55
6.1	Discussão	55
6.2	Trabalho futuro	56
	Bibliografia	57
	Anexo	63

Lista de Tabelas

2.1	Exemplo <i>TreeTagger</i>	8
2.2	Exemplo anotador de frases	8
4.1	Função SISTEMADEEXTRACAOODETERMOSRELEVANTES <i>input/output</i> .	15
4.2	Função TRATAMENTODETEXTO <i>input/output</i>	16
4.3	Função OCORRENCIASRELEVANTES <i>input/output</i>	18
4.4	Função FRASESRELEVANTES <i>input/output</i>	19
4.5	Função TREETAGRELATORIOS <i>input/output</i>	20
4.6	Função IDENTIFICARCOMPLEMENTOSDELUGAR <i>input/output</i>	23
4.7	Função IDENTIFICARCONDICOESCLINICAS <i>input/output</i>	24
4.8	Função IDENTIFICARDESCRICOESECONDICOESCLINICAS <i>input/output</i>	26
4.9	Função IDENTIFICAREXCECOES <i>input/output</i>	27
4.10	Função IDENTIFICARENTIDADES <i>input/output</i>	28
4.11	Função NUMERARENTIDADES <i>input/output</i>	32
4.12	Função OBTERSEQUENCIA <i>input/output</i>	33
4.13	Função ULTIMASEQUENCIA <i>input/output</i>	34
4.14	Função EXTRAIRCONDICAOCLINICA <i>input/output</i>	36
4.15	Função ELEMENARVERBOSECLUGAR <i>input/output</i>	37
4.16	Função EXTRACAODEDESCRICAODECONDICOESCLINICAS <i>input/output</i>	39
4.17	Função ELEMENARVERBOSECCCLINICASEDCCLINICAS <i>input/output</i>	39
4.18	Função EXTRACAODEDECCDELUGAR <i>input/output</i>	41

4.19	Função EXTRAIRTERMOSRELEVANTES <i>input/output</i>	43
5.1	Conjunto de dados exemplificativos	47
5.2	Resultados de Avaliação	48
5.3	Resultados da extração automática condições clínicas, primeiro conjunto	49
5.4	Resultados da extração automática de complementos circunstanciais de lugar, primeiro conjunto	50
5.5	Resultados da extração automática condições clínicas, segundo conjunto	52
5.6	Resultados da extração automática de complementos circunstanciais de lugar, primeiro conjunto	52

Lista de Figuras

1.1	Sistema de extração de termos relevantes	3
4.1	Exemplo extração de termo relevante	14
4.2	Exemplo complemento direto/condição clínica	30
4.3	Número de ocorrências verbo condição clínica ("nome")	30
4.4	Entidades associadas a uma ocorrência relevante	33
4.5	Entidades associadas a uma ocorrência relevante	34
4.6	Extração automática de condição clínica, uma frase	35
4.7	Extração automática de condição clínica, duas frases	35
4.8	Extração automática de descrições clínicas, exemplo 1	37
4.9	Extração automática de descrições clínicas, exemplo 2	37
4.10	Extração automática de descrições clínicas, exemplo 3	37
4.11	Extração automática de descrições clínicas, exemplo 4	38
4.12	Extração automática de complementos circunstanciais de lugar, exemplo 1	39
4.13	Extração automática de complementos circunstanciais de lugar, exemplo 2	39
4.14	Extração automática de complementos circunstanciais de lugar, exemplo 3	40
4.15	Extração automática de complementos circunstanciais de lugar, exemplo 4	40
4.16	Extração automática de complementos circunstanciais de lugar, exemplo 5	40
5.1	Identificação de entidades uma frase relevante	46
5.2	Identificação de entidades duas frases relevantes	46
5.3	Identificação de entidades frase 3	51

5.4	Identificação de entidades frase 4	51
5.5	Identificação de entidades frase 5	53
A.1	Palavras da família “nódulo”	63
A.2	Palavras da família “quisto” e “cisto”	64
A.3	Palavras da família “cóloide”	64
A.4	Palavras da família “sólido”	64
A.5	Palavras da família “misto”	64
A.6	Palavras da família “vascularização”	64
A.7	Palavras da família “calcificação”	65
A.8	Palavras da família “ecóicos”	65
A.9	Palavras da família “écogenico”	66
A.10	Palavras da família “quístico”/“cístico”	66

Lista de Blocos de Código

4.1	Sistema de extração de termos relevantes	15
4.2	Tratamento de texto	16
4.3	Identificação de ocorrências relevantes	17
4.4	Identificação de frases relevantes	19
4.5	<i>TreeTagger</i> em frases relevantes	20
4.6	Identificar complementos circunstanciais de lugar	22
4.7	Identificar condições clínicas	24
4.8	Identificar descrições de condições clínicas	25
4.9	Identificar exceções	27
4.10	Identificação de entidades	28
4.11	Identificar ocorrências relevantes	29
4.12	Identificar fim de frase	29
4.13	Identificação de verbos	30
4.14	Numerar entidades	31
4.15	Obter sequência	32
4.16	Obter última sequência	34
4.17	Extrair condição clínica	36
4.18	Eliminar verbos e complementos circunstanciais de lugar	36
4.19	Extração de descrições de condições clínicas	38
4.20	Eliminar verbos, condições clínicas e descrição de condições clínicas	39
4.21	Extração de complementos circunstanciais de lugar	41
4.22	Extrair termos relevantes	42
4.23	Extrair ocorrência relevante de sequência	42
4.24	Obter texto	43

Acrónimos

RE Reconhecimento de Entidades

PLN Processador de Linguagem Natural

EI Extração de Informação

RI Recuperação de Informação

POSTagging Part-of-speech tagging

Capítulo 1

Introdução

O principal objetivo deste trabalho de dissertação consiste no desenvolvimento de um sistema capaz de extrair termos clínicos relevantes de relatórios da tireoide. Para que tal aconteça, o sistema deverá ser capaz de extrair termos relevantes de relatórios clínicos escritos em português.

Neste capítulo da dissertação é feita uma exposição do assunto em estudo, da motivação e dos objetivos.

1.1 Contexto

Os relatórios clínicos são uma das principais ferramentas na gestão clínica e funcionam como um suporte informativo do paciente, onde os médicos registam o estado do paciente e os tratamentos que o paciente receberá no futuro [1].

Muita dessa informação que consta nos relatórios clínicos está pouco ou nada estruturada, pois na maior parte das vezes é apresentada na narrativa sob a forma de linguagem natural [2]. Guardar informação desta forma é conveniente para descrever conceitos e eventos, no entanto, torna difícil tarefas como pesquisas e análises estatísticas.

Todos estes entraves da linguagem natural tornam difícil o processo de gestão de informação médica. Todavia, a automatização de tarefas pode facilitar o processo. Um exemplo é o da extração automática de termos relevantes em relatórios clínicos.

A extração de termos relevantes começou fora do domínio clínico, sendo que muito do trabalho foi feito numa conferência patrocinada pelo governo dos Estados Unidos da América, *Message Understanding Conference* (MUC), entre 1987 e 1998 [3]. A MUC tinha como objetivo principal avaliar os diferentes sistemas de extração de termos relevantes em diferentes áreas.

Dentro da área médica, existem dois tipos de texto: o texto biomédico e o texto clínico. Podemos definir o texto biomédico como o texto que aparece, por exemplo, em artigos, livros, posters, etc. Já o texto clínico descreve os pacientes, as suas patologias, historial clínico, etc. Ao

processamento da linguagem clínica e biomédica chamamos processamento de linguagem médica.

O primeiro grande projeto de processamento de linguagem médica foi o *Linguistic String Project-Medical Language Processor* (LSP-MLP) [4]. O projeto envolvia quatro processos: *parsing*, decomposição sintática, regularização e mapeamento de informação numa tabela. Com base no LSP-MLP surge o *Medical Language Extraction and Encoding System* (MedLEE) [5], primeiro sistema na área clínica a extrair termos relevantes.

Um sistema de extração de termos relevantes passa por dois processos: o processamento de linguagem médica e o reconhecimento de entidades (RE). O processamento de linguagem médica tem como base Processadores de Linguagem Natural (PLN), onde se converte o texto desestruturado dos relatórios clínicos em texto estruturado. Só depois, é possível começar o RE, ou seja, o reconhecimento de palavras e frases médicas que representam os conceitos específicos do domínio de estudo.

1.2 Motivação

A tiroide é uma estrutura que se localiza em frente à traqueia ao nível das vértebras C5-T1, sendo uma das maiores glândulas endócrinas do corpo humano. Consiste principalmente em dois lobos, direito e esquerdo, e num istmo que une os dois lobos. É responsável por controlar o metabolismo e a taxa de calcitonina, uma hormona que controla o metabolismo do cálcio, e também afeta todas as áreas do corpo, exceto a própria, o baço, testículos e útero [6].

As causas das doenças associadas à tiroide dependem de inúmeros fatores, sendo que a causa mais comum de doenças na tiroide são as zonas não iodizadas: a deficiência de iodo pode levar à formação de bócios e hipotireoidismo. Já nas áreas iodizadas, as doenças podem variar entre hipotireoidismo e hipertireoidismo [7].

Os nódulos da tiroide são um exemplo de um problema clínico comum e estudos mostram que a prevalência de nódulos na tiroide é de aproximadamente 5% nas mulheres e 1% nos homens, para populações que vivem fora de zonas iodizadas. No entanto, a importância clínica dos nódulos tireóideos é saber se estes se podem traduzir em cancro da tiroide, o que ocorre entre 7% a 15% dos casos, dependendo da idade, sexo, história de exposição à radiação, história familiar e outros fatores [8].

Um estudo prevê que até 2019 o cancro da tiroide será o terceiro tipo de cancro mais comum entre mulheres norte americanas, resultando num custo que pode variar entre 19 e 21 bilhões de dólares para os Estados Unidos [9].

Dadas as circunstâncias, é inevitável que a quantidade de relatórios da tiroide aumente, tornando-se cada vez mais difícil arquivar e gerir a informação neles contida, ou seja, como os relatórios são escritos na forma de texto não estruturado, a consulta da informação relevante sobre cada paciente será cada vez mais difícil de ser consultada. A definição de uma metodologia que permita automatizar a compreensão de parte do conteúdo destes relatórios iria permitir

o apoio à decisão clínica, poupando recursos humanos de elevado custo e proporcionando um serviço de mais qualidade.

1.3 Objetivo

O objetivo deste projeto é o desenvolvimento de um sistema que seja capaz de extrair termos clínicos relevantes de um relatório, sendo que as entidades que constituem os termos relevantes a serem extraídos são: condições clínicas da tiroide, descrição das condições clínicas, dimensões das condições clínicas e zonas de incidência das condições clínicas.

O sistema desenvolvido pretende automatizar a compreensão dos relatórios clínicos, funcionando como um sistema de apoio à decisão clínica.

Para que tal seja possível é necessário que o sistema, tal como representado na Figura 1.1, seja capaz de converter a informação desestruturada dos relatórios em informação estruturada, reconhecer/identificar entidades e reconhecer/identificar ocorrências relevantes, ocorrências que estão associadas aos termos relevantes. Por fim, associar às ocorrências relevantes identificadas as diferentes entidades, extraindo desta forma os termos relevantes.

Estão disponíveis para análise um total de 1690 relatórios da tiroide. Os relatórios e os termos a serem extraídos foram fornecidos pela *FlipKick*.

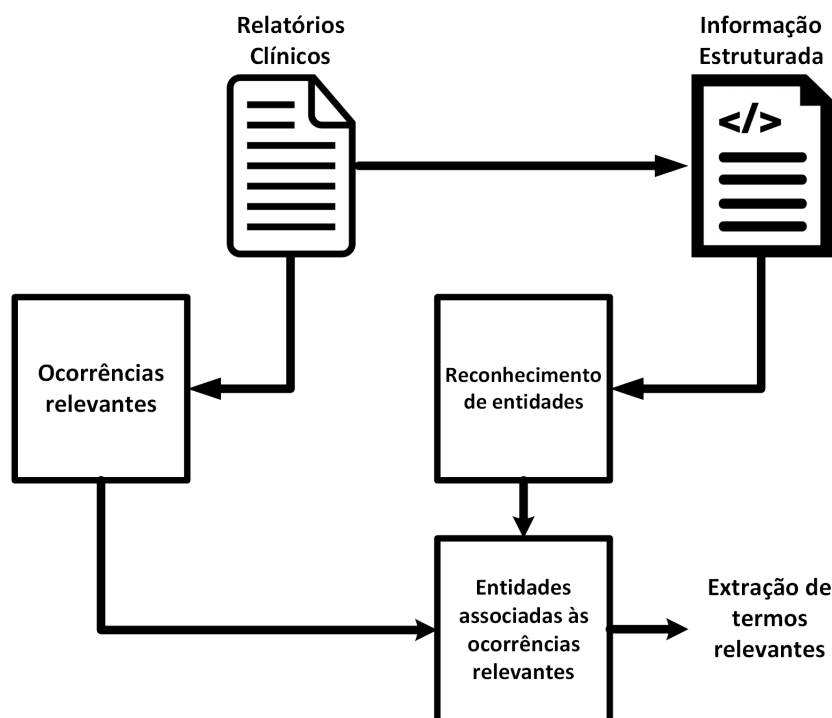


Figura 1.1: Sistema de extração de termos relevantes

1.4 Contribuições

A presente dissertação apresenta um sistema para extração automática de condições clínicas, descrição de condições clínicas, zonas de incidência das condições clínicas e dimensões das condições clínicas em relatórios da tiroide, escritos em português. A metodologia utilizada na extração de termos é pioneira para a língua portuguesa, sendo que a implementação do sistema trouxe algumas contribuições inovadoras. Sendo a principal, a construção de um novo modelo de representação da informação contida nos relatórios.

1.5 Estrutura da dissertação

A dissertação está organizada da seguinte forma:

Capítulo 2: ferramentas e conceitos utilizados. Neste capítulo da dissertação são apresentados vários conceitos e ferramentas usadas na solução proposta.

Capítulo 3: estado da arte. Neste capítulo, pretende-se abordar os progressos feitos na área da extração de termos relevantes em relatórios clínicos, tanto na língua inglesa como na língua portuguesa.

Capítulo 4: solução proposta. Neste capítulo será descrito o processo utilizado para atingir os objetivos propostos.

Capítulo 5: avaliação. Neste capítulo é apresentada a avaliação do sistema desenvolvido.

Capítulo 6: conclusões. Neste capítulo são apresentadas as conclusões.

Capítulo 2

Ferramentas e conceitos utilizados

Neste capítulo da dissertação são apresentados vários conceitos e ferramentas usadas na solução proposta. Primeiramente, iremos abordar conceitos relacionados com o processamento dos relatórios clínicos, como o “*Text Mining*”, o “Processamento de linguagem natural”, a “Extração de informação” e o “Reconhecimento de Entidades”. De seguida, apresentaremos os conceitos relevantes da dissertação e as ferramentas utilizadas.

2.1 *Text Mining*

Os relatórios clínicos contêm informação desestruturada. Sendo assim, para se realizar a extração automática dos termos relevantes, é necessário entender computacionalmente o que esta significa. Para o fazer, utilizamos técnicas de *Text Mining*.

Text Mining é definido como “O processo de extração de padrões e conhecimento, a partir de texto não estruturado” [10]. Como é mencionado em [10] o “*text mining* utiliza técnicas de recuperação de informação, extração de informação, processamento de linguagem natural (PLN), *data mining*, algoritmos de aprendizagem e estatística”.

2.2 Processamento de Linguagem Natural

De forma a entender computacionalmente o que significa a informação dos relatórios clínicos, é necessário fazer o processamento de linguagem natural que é definido como o “processo de converter linguagem natural numa representação formal que seja computacionalmente fácil de manipular” [11]. Um PLN pode ser visto como “uma ferramenta para o aperfeiçoamento de processos de extração de termos” [12].

Para extrair automaticamente termos clínicos recorreremos a tarefas importantes de PLN, como a análise morfológica, a análise sintática, a análise semântica e análise lexical. Estas tarefas facilitam o processo de reconhecimento de entidades, que por sua vez facilita a extração

automática de termos clínicos em relatórios.

Na presente dissertação são usados dois tipos de análises: a sintática e morfológica.

Na análise morfológica temos o *Part-of-speech tagging* (*POSTagging*) e a lematização. O *POSTagging* “tenta rotular/marcar cada palavra com a sua categoria gramatical” [13] (nome, verbo, adjetivo, etc.). Já a lematização, determina o lema de cada palavra, forma da palavra encontrada no dicionário. Por exemplo, o lema das palavras “nodulares”, “nodularidade” e “nódulos” deve ser “nódulo”.

Na análise sintática temos: o *parsing* e o anotador de frases, “que deteta as fronteiras de uma frase” [14], ou seja, o início e o fim de uma frase.

2.3 Extração de informação

A extração de termos relevantes de relatórios clínicos é uma tarefa de extração de informação (EI), que, como referido acima, é uma técnica de *Text Mining* e é definida em [15] como a “extração automática de tipos predefinidos de informação”. Esta não deve ser confundida com a recuperação de informação (RI), que é focada em selecionar documentos que contêm um determinado termo, como por exemplo, o motor de busca *Google* [16]. EI é focado em retornar informação ou factos, enquanto que a RI devolve documentos de uma coleção de documentos.

2.4 Reconhecimento de entidades

O reconhecimento de entidades é definido em [17] com uma “importante tarefa de EI” que tem como objetivo “identificar referências a entidades no texto”. Na presente dissertação podemos considerar que as entidades são os termos referentes às condições clínicas, descrições de condições clínicas e zonas de incidência das condições clínicas.

O RE pode ser separado em duas grandes classes: as abordagens baseadas em regras e as abordagens estatísticas. Abordagens baseadas em regras são feitas com base em condições/padrões de texto, como por exemplo, identificar o nome de uma pessoa. Imaginemos que se pretende identificá-lo nas seguintes frases: “Um grande exemplo de fé, em Portugal, é a devoção ao Dr. José Sousa Martins” e “Sr. Fernando Pinto de Sousa mora na freguesia da Prelada”. Nas frases antecedentes sabemos que o nome da pessoa se encontra logo a seguir aos pronomes de tratamento (Dr. e Sr.). Por outro lado, na abordagem estatística temos um processo de aprendizagem numérica. Voltemos à frase “Sr. Fernando Pinto de Sousa mora na freguesia da Prelada”, na qual se pretende identificar o nome da pessoa e o local. Para tal, pode-se recorrer a um modelo estatístico que diz quais são as palavras mais prováveis de pertencer à classe nome da pessoa (NM) e a classe local (LC), bem como as palavras que não pertencem a nenhuma das classes (NA). A frase acima identificada por classes, onde cada palavra representa uma classe, ficaria da seguinte forma “NA NM NM NA NM, NA NA NA NA LC” [18].

O reconhecimento de entidades, neste trabalho, é feito todo ele com abordagens baseadas em regras, pois as entidades são facilmente identificadas, não sendo necessário recorrer a métodos estatísticos.

2.5 Ocorrências relevantes

Para o caso em estudo, considera-se um termo relevante o conjunto de entidade(s) ou palavra(s) que descrevem uma condição clínica da tiroide. A cada termo relevante está associada uma ocorrência relevante.

Para identificar ocorrências relevantes em relatórios da tiroide partimos do pressuposto que estas se encontram em milímetros (“mm”), porque normalmente quando estamos perante esta ocorrência, esta corresponde a uma medida de uma determinada condição clínica, ou à medida máxima de uma mesma condição. Consequentemente, a uma condição clínica estão associadas entidades como a descrição da condição e a sua zona de incidência. Consideramos então que as entidades associadas a uma ocorrência relevante são: a condição clínica associada à mesma e as entidades associadas a essa condição.

2.6 *TreeTagger*

Para fazer o *POSTagging*, *parsing* e análise morfológica foi utilizada a ferramenta *TreeTagger* [19] com o *parameter file*: “*tree-tagger-portuguese-finegrained*” [20]. O *TreeTagger* fornece para cada uma das palavras a classe gramatical e o lema, forma da palavra encontrada no dicionário. Por exemplo, o lema do verbo “comerá” é “comer”, o lema do adjetivo “bonitas” é “bonito”, etc.

Para a frase “Na vertente esquerda do istmo há um nódulo sólido com áreas císticas centrais medindo 10x11mm.”, o *TreeTagger* produz um *output* igual ao apresentado na Tabela 2.1.

palavra	tag	lema
Na	SP+DA	em+a
vertente	NCFS000	vertente
esquerda	AQ0FS0	esquerdo
de	SPS00	de
o	DA0MS0	o
istmo	NCMS000	istmo
há	VMIP3S0	haver
um	DI0MS0	um
nódulo	NCMS000	nódulo
sólido	AQ0MS0	sólido
com	SPS00	com
áreas	NCFP000	área
císticas	AQ0FP0	cístico
centrais	NCFP000	central
medindo	VMG0000	medir
10x11mm	NCMS000	<unknown>
.	Fp	.

Tabela 2.1: Exemplo *TreeTagger*

Como podemos ver no exemplo referido acima, cada palavra apresenta uma etiquetagem. Por exemplo, “Em” é identificado como a contração da preposição (“SP”) “em” com o determinante artigo (DA) “a” e “há” é identificado como “VMIP3S0”, ou seja, um verbo “V”, principal “M”, indicativo “I”, na 3^o pessoa (P) do singular “3S”.

2.7 *Package openNLP*

Para anotar as frases dos relatórios foi utilizado o *package* “openNLP” do R [21] com o modelo “openNLPmodels.pt”, treinado com o corpus “bosque” [22]. Para o texto “No lobo esquerdo observa-se um nódulo calcificado com cerca de 8mm, de natureza residual. Não se definem imagens de outras lesões nodulares individualizáveis dominantes sólidas ou císticas.” é dado o *output* da Tabela 2.2.

id	type	start	end
1	sentence	1	89
2	sentence	91	189

Tabela 2.2: Exemplo anotador de frases

São identificadas duas frases, sendo que uma começa na posição “1” da *string* dada e acaba na posição “89” e outra começa na posição “91” e termina na “189”.

Capítulo 3

Estado da Arte

Neste capítulo, pretende-se abordar os progressos feitos na área da extração de termos relevantes em relatórios clínicos, tanto na língua inglesa como na língua portuguesa. Primariamente, abordaremos os progressos feitos no campo do PLN, de seguida, o RE e, por fim, alguns trabalhos de extração de termos relevantes.

3.1 Extração de termos relevantes

A extração de termos relevantes no domínio clínico não é uma novidade. Esta foi iniciada em 1995 [5] e tem tido utilidade em diversas áreas da medicina, como por exemplo, em diagnósticos [23], medicação [24] e radiologia [25]. A generalidade dos trabalhos realizados nesta área encontra-se na língua inglesa, e, apesar de existirem alguns trabalhos em português europeu relativamente à extração de termos, no domínio clínico pouco foi feito. No entanto, apesar da maioria dos estudos se encontrar em Inglês, estes têm um papel importante, pois em alguns casos podem ser aplicados e adaptados à língua portuguesa.

Um sistema de extração de termos relevantes passa por dois processos: o processamento de linguagem médica e o RE, existindo várias técnicas para ambos.

3.1.1 Processamento de linguagem médica

O texto clínico, dada a sua forma desestruturada, não pode ser imediatamente analisado pelo sistema de extração de termos relevantes. É necessário o processamento do texto clínico, ou seja, tornar o *input* desestruturado em *input* estruturado para ser analisado posteriormente pelo sistema de extração de termos relevantes. No processamento de texto clínico podemos destacar os seguintes métodos: corretor ortográfico [26], eliminar ruído [27], desambiguação de palavras ambíguas [28], rotular palavras ou expressões [29] e representação de conteúdo [30].

1. Imaginemos o seguinte caso: “Cancro” e “Cncro”. O sistema de extração de termos

relevantes identifica “Cancro” e “Cncro” como palavras distintas, embora “Cncro” esteja mal escrito. Para que ambas sejam identificadas como a mesma palavra, é fundamental o uso de corretores ortográficos. Existem alguns na língua inglesa dedicados ao texto clínico, como o *UMLS-based spell checker* [26].

2. Nos relatórios clínicos, existe um grande número de palavras desconhecidas do ponto de vista do processador de linguagem médica, nomeadamente, vocabulário médico, acrónimos, abreviaturas, etc. Deste modo, é importante que as mesmas sejam tratadas, eliminando o ruído e evitando más análises por parte do PLN. Isto é feito em [27].
3. Desambiguação de palavras ambíguas, ou seja, entender o contexto destas em determinadas situações, por exemplo, palavras que possam ser substituídas por outras equivalentes (sinónimos), homógrafas, homónimas ou parónimas [28].
4. Rotular palavras ou expressões no contexto clínico, ou seja, um *POSTagging* aperfeiçoado para o contexto clínico [29]. As más anotações poderão afetar a extração de termos relevantes no texto clínico.

3.1.2 Reconhecimento de entidades

Depois de se processar o texto clínico, pode-se começar a extração automática dos termos relevantes. Um passo fundamental na extração é o RE, pois estas representam os conceitos específicos do domínio de estudo. A extração do termo é a abordagem mais simples e consiste na extração automática de todas as ocorrências de um termo. Imaginando que o termo pretendido é “cancro”, ter-se-á de extrair do texto clínico todas as ocorrências de “cancro”. Já a extração de expressão é uma abordagem em tudo semelhante à extração de termo só que, em vez de um termo, seleciona-se uma expressão, por exemplo, todas as ocorrências de “cancro maligno”. Ambas poderão ser úteis para localizar termos específicos, mas pouco úteis em situações ambíguas, revelando-se pouco poderosas na extração de termos relevantes [31]. Para as situações ambíguas é necessário fazer o RE. Para o fazer, existem várias abordagens, sendo elas: usar vocabulário médico para detetar palavras relacionadas com patologias [32], lidar com negações [33], compreender as relações temporais [34] e atribuir códigos de doença aos relatórios clínicos [35].

1. Usar vocabulário médico para detetar palavras relacionadas com patologias e, desta forma, consegue-se distinguir termos que se referem a termos que não se referem a patologias. Em [32] é criada uma ontologia, modelo de dados com as condições clínicas de determinado domínio, para as doenças pulmonares.
2. Lidar com negações [33]. Muitas vezes deparamo-nos com frases na negativa em relatórios clínicos, o que poderá representar um problema para a extração automática de termos relevantes. Imaginemos o seguinte caso em que se pretende selecionar todos os relatórios referentes à patologia “nódulos” e existem relatórios com as seguintes frases: “O paciente apresenta nódulos” e “O paciente não apresenta nódulos”. O primeiro relatório refere-se, de

facto, a nódulos, já o segundo nada terá a ver com “nódulos”, pois, como é dito, o paciente não apresenta a patologia.

3. Compreender as relações temporais [34], ou seja, compreender a sequência de eventos clínicos ao longo do tempo e extrair o evento relevante.
4. Atribuir um código ICD (*International Statistical Classification of Diseases and Related Health Problems*) a relatórios clínicos. Para atribuir um código ICD é necessário extrair automaticamente dos relatórios as condições clínicas abordadas nos mesmos [35]. Os ICD's classificam cada uma das doenças com um código. Cada código pertence a uma categoria, sendo as categorias constituídas por códigos semelhantes, ou seja, doenças semelhantes têm códigos semelhantes. O ICD serve para efeitos estatísticos, por exemplo, para identificar um surto de determinada doença. Em Portugal, o ICD é utilizado em todos os hospitais do Serviço Nacional de Saúde [36].

3.1.3 Sistemas de extração de termos relevantes

Quanto aos trabalhos de extração automática de termos relevantes na área clínica, podemos destacar os seguintes:

Na língua inglesa, dado o seu uso massivo e pioneirismo, o *The Linguistic String Project–Medical Language Processor* (LSP–MLP) [37] e o *Medical Language Extraction and Encoding system* (MedLEE) [5].

1. O *The Linguistic String Project–Medical Language Processor* (LSP–MLP) [37], sob a liderança de Sager, foi pioneiro no desenvolvimento de um sistema de processamento de linguagem médica. Este foi aplicado a uma variedade de domínios, como cartas de alta, radiologia, asma, artrite reumatóide e farmacologia. Este sistema produz uma saída estruturada onde os termos são padronizados, mas não codificados e possui fontes de conhecimento separadas, que especificam propriedades sintáticas e semânticas.
2. Por outro lado, temos o MedLEE que foi projetado, primariamente, no centro médico de *Columbia-Presbyterian* em relatórios de radiologia e posteriormente foi utilizado noutros domínios. Trata-se de um sistema de apoio à decisão para extração automática de informação relevante em relatórios clínicos, sendo responsável por estruturar e codificar a informação clínica para que os dados possam ser usados, posteriormente, em processos automatizados [5].

Além disso, podemos destacar outros sistemas da língua inglesa, de menor importância, tais como: *Special Purpose Radiology Understanding System* (SPRUS) [25]; *Symbolic Text Processor* (SymText) [38]; *The Mayo clinical Text Analysis and Knowledge Extraction System* (cTAKES) [39] e *REgenstrief eXtraction tool* (REX) [40].

1. O SPRUS é um sistema de extração de termos baseado na semântica do texto. Esta ferramenta usa informações semânticas de um sistema especializado em analisar relatórios de radiologia para fazer o *parse* do texto clínico e, desta forma, extrai e codifica as interpretações dos radiologistas. As interpretações codificadas são armazenadas numa base de dados clínicos para posteriormente o sistema reconhecer automaticamente as interpretações dos radiologistas. [25].
2. *SymText*, posterior ao SPRUS, combina um analisador sintático que usa “*augmented transition networks*” e análise semântica probabilística baseada em redes *Bayesianas*. Um dos principais objetivos desta pesquisa é explorar as interações entre a sintaxe e a semântica [38].
3. cTAKES é um sistema que usa PLN *open-source* (*openNLP* [41]) para a EI de registos clínicos. O cTakes é responsável por anotar o texto clínico, para posteriormente, estas anotações serem usadas por métodos e módulos, para processamento semântico de nível superior [39].
4. REX sistema de processamento de linguagem natural para extrair e codificar dados clínicos de relatórios. O sistema foi projetado para ser facilmente modificado e adaptado a diferentes tipos de relatórios clínicos: relatórios de radiologia, registos clínicos e cartas de alta [40].

Alguns dos trabalhos realizados na língua inglesa recorrem ao Sistema de Linguagem Médica Unificada (UMLS) [42], que funciona como uma fonte de conhecimento externo e é constituído por ontologias, ou seja, vocabulário estruturado de conceitos médicos e as suas relações. Desta forma, é possível compreender se estamos perante termos médicos (doenças, sintomas ou diagnósticos). O UMLS possui três fontes de conhecimento: a *Metathesaurus* (base de dados de vocabulário médico que conta com uma versão em português), a *Semantic Network* (estabelece relações entre o vocabulário médico do *Metathesaurus*) e a *SPECIALIST Lexicon* (léxico de termos biomédicos da língua inglesa). O UMLS conta, também, com a ferramenta *Lexical Tools* que é capaz de detetar variantes ortográficas, sinónimos, variantes de doenças, tempos verbais, entre outras [43].

No português europeu, podemos destacar apenas o *Medical Information eXtraction* (MedInX). O MedInX reconhece automaticamente, com base na extração de termos relevantes de texto clínico desestruturado, a doença abordada em cada relatório clínico atribuindo-lhe um código ICD. Usa PLN e no reconhecimento de termos relevantes lida com negações e recorre ao UMLS como fonte de conhecimento externo. A avaliação do sistema é de 95% de *precision* e de *recall* [44].

Capítulo 4

Solução proposta

Neste capítulo descrevemos o método, baseado em padrões, utilizado para atingir os objetivos propostos.

Um termo relevante é constituído por uma ocorrência relevante, dimensão de determinada condição clínica, e por outras entidades (zona de incidência da condição clínica, condição clínica e relativa descrição). Para a frase “Identifica-se um nódulo sólido com 14mm, no lobo esquerdo.” identificamos a condição clínica “nódulo”, a descrição da condição clínica como “sólido”, a dimensão da condição clínica (ocorrência relevante) como “14mm” e a zona de incidência da condição clínica como “no lobo esquerdo”. O termo relevante para a condição clínica “nódulo” mencionada na frase anterior, deverá ser “nódulo sólido 14mm no lobo esquerdo”. Para todos os termos relevantes temos uma ocorrência relevante, ou seja, uma dimensão de determinada condição clínica.

Partindo desse pressuposto, o sistema deverá extrair termos relevantes com base na ocorrência relevante e, para isso, teremos de a identificar. Depois de identificada a ocorrência, teremos de identificar as restantes partes do termo relevante, ou seja, as restantes entidades que pertencem ao termo relevante. Para uma ocorrência relevante existem várias entidades associadas, entidades que pertencem ao mesmo termo relevante da ocorrência. Para entender quais as entidades associadas às ocorrências relevantes é gerada uma sequência que reflete a ordem de aparição das diferentes entidades, como exemplificado na Figura 4.1, onde “c_lugar” é uma zona de incidência de uma condição clínica, “num” uma ocorrência relevante/dimensão da condição clínica, “adj” uma descrição de condição clínica e “nome” a condição clínica.


```

function SISTEMADEEXTRACAODETERMOSRELEVANTES(relatorio)
  frasesRelevantes ← FRASESRELEVANTES(relatorio)
  taggedFrasesRelevantes ← IDENTIFICARENTIDADES(frasesRelevantes)
  termosRelevantes ← EXTRAIRTERMOSRELEVANTES(taggedFrasesRelevantes)
  return termosRelevantes

```

Bloco de Código 4.1: Sistema de extração de termos relevantes

<i>Input</i>	<i>Output</i>			
	OR	cClínica	dCClínica	zClínica
"ECOGRAFIA DA TIRÓIDE Técnica Foram realizados ecotomogramas sagitais e transversais da glândula tiróide. Relatório: A glândula da tiróide está ligeiramente aumentada, medindo o lobo direito 59x22x15mm, o lobo esquerdo 53x23x13mm, apresentando o istmo uma espessura de 3mm. A ecoestrutura glandular é homogênea, identificando-se na porção média do lobo direito uma imagem nodular hipoecogénica com 6mm. Ainda no lobo direito observam-se alguns pequenos quistos não excedendo 5mm."	3mm			
	6mm	"nodular"	"hipoecogénica"	"na porção média do lobo direito"
	5mm	"quistos"		"no lobo direito"

Tabela 4.1: Função SISTEMADEEXTRACAOODETERMOSRELEVANTES *input/output*

O sistema de extração deve produzir o resultado da Tabela 4.1, onde “OR” são as ocorrências relevantes do relatório dado no *input*, “cClínica” as condições clínicas, “dCClínica” as descrições das condições clínicas e “zClínica” a zona de incidência das condições clínicas.

4.1 Pré-processamento de texto clínico

Nos relatórios é preciso garantir que o texto clínico esteja bem formatado, de forma a evitar más interpretações por parte do *TreeTagger*, ferramenta utilizada para identificar a classe gramatical e lema de cada palavra. O *TreeTagger* identifica cada uma das palavras com base no espaçamento, por isso, situações como espaçamento entre palavras inadequado ou mau uso de maiúsculas/minúsculas deverá ser corrigido primeiramente, de forma a garantir que o *TreeTagger* faça uma análise correta.

Desta forma, procedemos a algumas alterações no texto, nomeadamente:

- Eliminação de caracteres de escape (ex. $\backslash n \backslash t$)
- Eliminação do espaçamento entre unidade (ex. “1 cm” para “1cm”)
- Eliminação do espaçamento em unidades multidimensionais (ex. “12 x 20 mm” para “12x20mm”)
- Substituição de “/” por “ ” (ex. “médio/superior” para “médio superior”), para o *TreeTagger* interpretar ambas as palavras
- Conversão de palavras em maiúsculas em minúsculas, quando usadas de forma irregular (Ex. “RELATÓRIO” para “relatório”)

function TRATAMENTODETEXTO(relatorio)

```

relatorio ← SUBSTITUIR("\[a - z]", "", relatorio)
relatorio ← SUBSTITUIR("([0 - 9]+) mm", "\1mm", relatorio)
relatorio ← SUBSTITUIR("([0 - 9]+) cm", "\1cm", relatorio)
relatorio ← SUBSTITUIR("([0 - 9]+) [m|M]hz", "\1\2", relatorio)
relatorio ← SUBSTITUIR("(\s+)([x|X])(\s+)", "\2", relatorio)
relatorio ← SUBSTITUIR("([a - z])(/)([a - z])", "\1 \3", relatorio)
relatorio ← TOLOWER(relatorio)
relatorio ← SUBSTITUIR("(\.)(\s+)([a - z])", ". \U\3", relatorio)
return relatorio

```

Bloco de Código 4.2: Tratamento de texto

A função “SUBSTITUIR” é uma função que substitui uma *substring*, que corresponde a uma expressão regular ou não, contida numa *string*, por outra *substring*. O primeiro parâmetro corresponde à *substring* a ser substituída, o segundo à *substring* de substituição e o terceiro à *string* que queremos alterar. Nas expressões regulares temos o “\” como o início de um carácter especial (ex. “\” é “\”, “+” é “+” e “#” é “#”), [0-9] para dígitos, [a-z] para letras de “a” a “z”, “+” ocorrer uma ou mais vezes, “[a|b]” “a” ou “b” e “s” como espaço. “\1”, “\2”, “\3”, etc. corresponde ao primeiro, segundo, terceiro, etc., conforme a numeração, conteúdo da *substring* a ser substituída, que está entre parêntesis, ou seja, “\1” corresponde ao primeiro conteúdo que se encontra entre parêntesis. “\U” converte em maiúscula o conteúdo que o precede (“\1”, “\2”, “\3”, etc.).

<i>Input</i>	<i>Output</i>
"Na transição do terço médio/inferior da HEMITIROIIDE direita persiste um nódulo quístico com 3 mm."	"Na transição do terço médio inferior da hemitiroide direita persiste um nódulo quístico com 3mm."

Tabela 4.2: Função TRATAMENTODETEXTO *input/output*

4.2 Identificação de possíveis ocorrências relevantes

Para identificar ocorrências relevantes partimos do pressuposto que estas, geralmente, são medidas que se encontram em milímetros (“mm”).

Para as identificar, primeiro, detetaram-se as palavras que começam com número e posteriormente descartaram-se os casos não relevantes. Para tal, eliminaram-se todos os números com medidas tridimensionais, embora terminem em “mm”, referem-se, apenas, a dimensões dos lobos ou do istmo (ex. “O lobo esquerdo mede respetivamente 35x11x11mm.”). Uma medida tridimensional é constituída pela sequência “*numero*” → “*x*” → “*numero*” → “*x*” → “*numero*”. Além disso, eliminaram-se situações em que a palavra começava por número mas não terminava na unidade milímetros ou apresentava demasiados algarismos para ser uma dimensão de uma condição clínica, como medidas em centímetros (“cm”), em mega-hertz (“MHz”) e datas (ex. “2cm”, “7.5mhz” e 2015).

Quanto aos números que não apresentam nenhuma unidade, fez-se uma análise individual para determinar aqueles que se encontram em enumerações. Se esta terminar em “mm” é relevante (ex. “No lobo esquerdo três hipocóicos com cerca 7, 9 e 10mm.”), se não, é irrelevante. Relativamente aos restantes números em “mm” todos eles são relevantes, e os em cm são irrelevantes.

Note-se que as ocorrências relevantes baseadas nos pressupostos acima referidos são apenas possíveis ocorrências relevantes, pois muitas delas podem nem se referir a condições clínicas (Ex. “Medindo o istmo cerca de 3mm.”). Só depois de entender quais as entidades associadas a uma ocorrência relevante é que é possível descartá-las.

```
function OCORRENCIASRELEVANTES(texto)
  palavras ← DIVIDIR((texto, " "))
  numeros ← OBTER(" [1 - 9]", palavras)
  aRemover ← OBTER(" ^ ([0 - 9]+)(x)([0 - 9]+)(x)([0 - 9]+)$", palavras)
  aRemover ← aRemover ∪ OBTER("[(19)|(20)]\d{2}", palavras)
  aRemover ← aRemover ∪ OBTER("mhz", palavras)
  numeros ← numeros \ aRemover
  relevantes ← {}
  inicio ← 1
  for all n ∈ numeros do
    if TEM("mm", n) then
      relevantes ← relevantes ∪ numeros[inicio : POSICAO(n)]
      inicio ← POSICAO(n) + 1
    else if TEM("cm", n) then
      inicio ← POSICAO(n) + 1
  return relevantes
```

Bloco de Código 4.3: Identificação de ocorrências relevantes

A função “DIVIDIR” é uma função que cria uma lista de *subtrigs* de uma *string* dividida consoante um padrão escolhido. O primeiro parâmetro corresponde à string a ser dividida e o segundo ao padrão que a divide. Se o padrão que divide for , para a frase “Um exemplo.” obtemos uma lista com os seguintes elementos: “Um” e “exemplo.”.

A função “OBTER” é uma função que seleciona elementos de uma lista segundo uma expressão regular. O primeiro parâmetro corresponde à expressão regular e o segundo à lista. “^” corresponde ao início da palavra, “\$” ao fim da palavra e “{2}” ao número de vezes exato que se pode repetir a expressão anterior. Com a expressão regular “^[1-9]”, por exemplo, obtemos todas as palavras que começam por dígito.

A função “TEM” é uma função que determina se uma *substring*, que corresponde a uma expressão regular ou não, está contida numa *string* ou não. Se esta estiver, retorna verdadeiro, se não, retorna falso.

A função “POSICAO” é uma função que retorna a posição na lista de um elemento que pertence a essa lista, sendo que o único parâmetro é um elemento que pertencente a uma lista.

<i>Input</i>	<i>Output</i>
"Foram efectuados cortes ecotomográficos sagitais e transversais para estudo da glândula tiróide, utilizando sonda linear de 7.5MHz. O lobo direito mede 45x15x14mm. No lobo esquerdo os três mais significativos medem cerca de 4, 7 e 11mm. No lobo direito há um microquisto com 2mm. Longitudinal 4cm, transversal 1,3cm, antero-posterior 1,2cm"	"4", "7", "11mm", "2mm"

Tabela 4.3: Função OCORRENCIASRELEVANTES *input/output*

4.3 Identificação de frases relevantes

De forma a tornar o processo mais eficiente foram consideradas para extração dos termos, apenas, frases relevantes, ou seja, todas aquelas que contêm uma ou mais ocorrências relevantes. Desta forma é possível evitar que partes do texto do relatório, ou até mesmo todo o relatório, tenham de ser interpretadas pelo *TreeTagger*.

As frases relevantes, na maior parte dos casos, contêm as entidades associadas às ocorrências relevantes, embora, em alguns casos, as entidades associadas ocorram na frase anterior. Deste modo, é imperativo selecionar a frase que contém a(s) ocorrência(s) relevante(s), bem como, a frase anterior.

```

function FRASESRELEVANTES(relatorio)
  texto ← TRATAMENTODETEXTO(relatorio)
  listaDeFrases ← OBTERFRASES(texto)
  frasesRelevantes ← {}
  for i = 1 to LENGTH(listaDeFrases) do
    if LENGTH(OCORRENCIASRELEVANTES(listaDeFrases[i])) > 0 then
      if i - 1 == 0 then
        frasesRelevantes ← frasesRelevantes ∪ listaDeFrases[i]
      else
        frasesRelevantes ← frasesRelevantes ∪ CONCAT(listaDeFrases[i-1], listaDeFrases[i])
  return frasesRelevantes

```

Bloco de Código 4.4: Identificação de frases relevantes

A função “OBTERFRASES” retorna uma lista de frases de um determinado texto, em que este é o único parâmetro, sendo o *package openNLP* responsável pela função.

A função “CONCAT” concatena numa *string* as variáveis que são dadas como parâmetros.

A função “LENGTH” é uma função que retorna o tamanho de determinada lista, sendo que o único parâmetro é uma lista.

<i>Input</i>	<i>Output</i>
"Na transição do terço médio com o terço inferior da hemitiroide direita persiste um nódulo quístico com 3mm. No istmo e na hemitiroide esquerda não se evidenciam nódulos, tanto de estrutura quística como de estrutura sólida. Não se registam alterações no espaço pré-traqueal ou na região dos vasos cervicais, nomeadamente adenomegalias. Alterações morfoestruturais difusas da glândula tiroide sugestivas de tiroidite crónica. Nódulo quístico com 3mm na hemitiroide direita."	"Na transição do terço médio com o terço inferior da hemitiroide direita persiste um nódulo quístico com 3mm.", "Alterações morfoestruturais difusas da glândula tiroide sugestivas de tiroidite crónica. Nódulo quístico com 3mm na hemitiroide direita."

Tabela 4.4: Função FRASESRELEVANTES *input/output*

4.4 *POSTagging* e lematização de frases relevantes

Para facilitar o reconhecimento de entidades utilizamos o *TreeTagger*. O *TreeTagger* é responsável pelo *POSTagging*, que identifica a classe gramatical (verbo, adjetivos, nomes, etc.), e por identificar o lema de cada palavra.

```

function TREETAGRELATORIOS(relatorio)
  frasesRelvantes  $\leftarrow$  FRASESRELEVANTES(relatorio)
  taggedRel  $\leftarrow$  {}
  for all  $f \in$  frasesRelevantes do
    taggedRel  $\leftarrow$  taggedRel  $\cup$  TREETAG( $f$ )
  return taggedRel

```

Bloco de Código 4.5: *TreeTagger* em frases relevantes

A função “*TREETAG*” é responsável pelo uso da ferramenta *TreeTagger*.

A variável “taggedRel” tem três atributos: palavra, tag e lema.

<i>Input</i>	<i>Output</i>		
	palavra	tag	lema
"Visualiza-se no lobo esquerdo pequeno quisto colóide com 4mm."	Visualiza	VMIP3S0	visualizar
	se	PP3CN000	se
	em	SPS00	em
	o	DA0MS0	o
	lobo	NCMS000	lobo
	esquerdo	AQ0MS0	esquerdo
	pequeno	AQ0MS0	pequeno
	quisto	AQ0MS0	quisto
	colóide	NCMS000	colóide
	com	SPS00	com
	4mm	NCMS000	<unknow>
	.	Fp	.

Tabela 4.5: Função TREETAGRELATORIOS *input/output*

4.5 Reconhecimento de entidades

Com o *POSTagging* e a identificação dos lemas de cada palavra torna-se, agora, muito mais fácil, identificar entidades como complementos circunstanciais de lugar, condições clínicas e descrições de condições clínicas.

4.5.1 Complemento Circunstancial de Lugar

O complemento circunstancial indica uma circunstância de facto expressa pelo verbo e pode ser de tempo, lugar, causa, etc., conforme a circunstância que se exprime na frase. Um complemento

circunstancial de lugar responde à pergunta “onde?” e pode surgir em qualquer ponto da frase, normalmente, entre vírgulas [45].

Em relatórios da tiroide, um complemento circunstancial de lugar é uma zona referente à tiroide, ou seja, as possíveis zonas de incidência das condições clínicas, onde se destacam as seguintes: lobo esquerdo, lobo direito e istmo [6]. Na frase “No istmo, há um nódulo sólido com áreas císticas centrais medindo 10x11mm.” podemos identificar o complemento circunstancial de lugar: “No istmo”.

Para identificar os complementos circunstanciais de lugar nos relatórios da tiroide tivemos em conta o seguinte: estes normalmente começam com a preposição “em” ou uma contração da preposição “em” (ex. no, na, num, numa...) e terminam nos lemas “istmo”, “ístmico”, “direito” ou “esquerdo”, ou seja, o término das zonas relevantes da tiroide. O lema “lobo” também foi considerado para os casos que se referem a ambos os lobos (ex. “em ambos os lobos”) ou para os casos que se referem a um lobo mas sem referência a qual (ex. “num dos lobos”). Determinar uma contração da preposição “em” é possível com o *TreeTagger* pois este identifica, por exemplo, “na” como “em+a”. Foram também considerados os casos “à direita” e “à esquerda” como relevantes.

O tamanho máximo considerado para um complemento circunstancial de lugar foi de 12 palavras, sem pontuação pelo meio (ex. vírgulas, pontos, etc.). Por exemplo, para a frase “A ecoestrutura do tecido glandular é levemente heterogénea visualizando-se na vertente anterior do terço médio do lobo esquerdo nódulo hipoecóico com 7x4mm.” é capaz de identificar o complemento circunstancial de lugar “na vertente anterior do terço médio do lobo esquerdo”.

```

function IDENTIFICARCOMPLEMENTOSDELUGAR(taggedRel)
  alvo ← 1
  i ← 1
  while i = 1 ≤ LENGTH(taggedRel) do
    if TEM("à", taggedRel[i]$palavra) and (TEM("esquerdo", taggedRel[i + 1]$lema) or
    TEM("direito", taggedRel[i + 1]$lema)) then
      taggedRel[i : (i + 1)]$entidade ← "c_lugar"
      i ← i + 2
      alvo ← i
    if TEM("em", taggedRel[i]$lema) then
      j ← i + 1
      while j ≤ LENGTH(taggedRel) do
        if TEM("lobo", taggedRel[j]$lema) then
          alvo ← j
        if TEM(lugares, taggedRel[j]$lema) then
          alvo ← j
          break
        j ++
    if alvo ≠ 1 and not(TEM("[: punct :]", taggedRel[i : alvo]$palavra)) and alvo - i ≤
12 then
      taggedRel[i : alvo]$entidade ← "c_lugar"
    else
      alvo ← i
      i ← alvo + 1
  return taggedRel

```

Bloco de Código 4.6: Identificar complementos circunstanciais de lugar

À variável “taggedRel” adicionamos o atributo “entidade”, para identificar as diferentes entidades.

“[: punct :]” é uma expressão regular que representa os caracteres de pontuação (“(, ”), “*”, “+”, “-“, “:”, etc.).

“z[x:y]” todas as posições de “x” a “y” da variável z.

A variável “lugares” é uma lista composta pelos seguintes lemas: “direito”, “esquerdo”, “istmo” e “ístmica”.

<i>Input</i>			<i>Output</i>			
palavra	tag	lema	palavra	tag	lema	entidade
Visualiza	VMIP3S0	visualizar	Visualiza	VMIP3S0	visualizar	
se	PP3CN000	se	se	PP3CN000	se	
em	SPS00	em	em	SPS00	em	c_lugar
o	DA0MS0	o	o	DA0MS0	o	c_lugar
lobo	NCMS000	lobo	lobo	NCMS000	lobo	c_lugar
esquerdo	AQ0MS0	esquerdo	esquerdo	AQ0MS0	esquerdo	c_lugar
pequeno	AQ0MS0	pequeno	pequeno	AQ0MS0	pequeno	
quisto	AQ0MS0	quisto	quisto	AQ0MS0	quisto	
colóide	NCMS000	colóide	colóide	NCMS000	colóide	
com	SPS00	com	com	SPS00	com	
4mm	NCMS000	<unknow>	4mm	NCMS000	<unknow>	
.	Fp	.	.	Fp	.	

Tabela 4.6: Função IDENTIFICARCOMPLEMENTOSDELUGAR *input/output*

4.5.2 Condições Clínicas

Embora o *TreeTagger* identifique a generalidade dos lemas, este, por vezes, falha. Por exemplo, para a palavra “pseudonodular”, o *TreeTagger* deveria retornar o lema “nódulo”. Em vez disso, este retorna “<unknow>”, ou seja, desconhece o lema para a palavra “pseudonodular”. Outra das falhas surge quando a determinada palavra está associado um erro ortográfico, por exemplo, “nodulo” (sem acento) o lema da palavra “nodulo” é identificado como “nodular”.

Para evitar erros, ao contrário dos complementos circunstanciais de lugar nos quais utilizamos os lemas, foi necessário determinar as variantes das condições clínicas. Para tal, tivemos em conta as patologias principais associadas à tiroide, tais como quisto/cisto, nódulo e bócio, bem como as palavras da família das patologias principais.

Para identificar as palavras da família “nódulo”, por exemplo, identificaram-se, em todos os relatórios, os casos positivos, palavras que pertencem à família de “nódulo”, e os casos falsos, palavras que não pertencem à família de “nódulo”. Para o fazer começamos por identificar o radical, constituinte da palavra indecomponível que se mantém depois de eliminar todos os afixos (ex. o radical das palavras “nódulo”, “nodularidade”, “nodulares” é “nodul”) [46], ou parte do radical e posteriormente verificamos se as palavras que contêm o radical ou parte dele são todas casos positivos. Se não o forem, podemos filtrar os casos negativos até conseguirmos um resultado com a totalidade de casos positivos. O radical escolhido, para os casos em que este tem acento, deve ser com acento e sem acento.

As palavras que contêm “nód” ou “nod” apresentam todas casos positivos para as palavras da família “nódulo” (ver Anexo A.1).

As palavras que contêm “bóc” ou “boc” apresentam todas casos positivos para as palavras da família “bócio”. Foi apenas identificada a palavra “bócio”.

As palavras que contêm “quisto” e “cisto” apresentam todas casos positivos para as palavras da família “quisto” e “cisto” (ver Anexo A.2).

```
function IDENTIFICARCONDICOESCLINICAS(taggedRel)
  for  $i = 1$  to LENGTH(taggedRel) do
    if taggedRel[i]$palavra  $\in$  condicoesClinicas then
      taggedRel[i]$entidade  $\leftarrow$  "nome"
  return taggedRel
```

Bloco de Código 4.7: Identificar condições clínicas

A variável “condicoesClinicas” é uma lista composta pelas palavras da família de "nódulo", "quisto/cisto" e "bócio", referidas acima.

<i>Input</i>			<i>Output</i>			
palavra	tag	lema	palavra	tag	lema	entidade
Visualiza	VMIP3S0	visualizar	Visualiza	VMIP3S0	visualizar	
se	PP3CN000	se	se	PP3CN000	se	
em	SPS00	em	em	SPS00	em	
o	DA0MS0	o	o	DA0MS0	o	
lobo	NCMS000	lobo	lobo	NCMS000	lobo	
esquerdo	AQ0MS0	esquerdo	esquerdo	AQ0MS0	esquerdo	
pequeno	AQ0MS0	pequeno	pequeno	AQ0MS0	pequeno	
quisto	AQ0MS0	quisto	quisto	AQ0MS0	quisto	nome
colóide	NCMS000	colóide	colóide	NCMS000	colóide	
com	SPS00	com	com	SPS00	com	
4mm	NCMS000	<unknow>	4mm	NCMS000	<unknow>	
.	Fp	.	.	Fp	.	

Tabela 4.7: Função IDENTIFICARCONDICOESCLINICAS *input/output*

4.5.3 Descrições de condições clínicas

Na descrição das condições clínicas, tal como nas condições clínicas, a identificação de lemas falha, ou seja, também é necessário identificar as variantes das descrições das condições clínicas. Estas são: “colóide”, “compressão”, “mergulhante”, “sólido”, “misto”, “halo”, “vascularização”, “calcificação”, “ecóico” e “ecogénico”, bem como as palavras da família das descrições mencionadas.

As palavras que contêm “coloi” ou “colói” apresentam todas casos positivos para as palavras da família “colóide” (ver Anexo A.3).

As palavras que contêm “compress” apresentam todas casos positivos para as palavras da família “compressão”. Foi apenas identificada a palavra “compressão”.

As palavras que contêm “mergulh” apresentam todas casos positivos para as palavras da família “mergulhante”. Foi apenas identificada a palavra “mergulhante”.

As palavras que começam por “sol” ou “sól” apresentam todas casos positivos para as palavras da família “sólido” (ver Anexo A.4).

As palavras que começam por “mis” apresentam todas casos positivos para as palavras da família “misto” (ver Anexo A.5).

As palavras que começam por “halo” apresentam todas casos positivos para as palavras da família “halo”. Foi apenas identificada a palavra “halo”.

As palavras que contêm “vascu” menos o caso negativo “cardiovascular”, apresentam todas casos positivos para as palavras da família “vascularização” (ver Anexo A.6).

As palavras que contêm “calci” apresentam todas casos positivos para as palavras da família “calcificação” (ver Anexo A.7).

As palavras que contêm “ecoi” ou “ecói” apresentam todas casos positivos para as palavras da família “ecóicos” (ver Anexo A.8).

As palavras que contêm “ecogé” ou “ecoge” apresentam todas casos positivos para as palavras da família “écogenico” (ver Anexo A.9).

```
function IDENTIFICARDESCRICOESDECONDICOESCLINICAS(taggedRel)
  for  $i = 1$  to LENGTH(taggedRel) do
    if taggedRel[ $i$ ]$palavra  $\in$  descricoesDecondicoesClinicas then
      taggedRel[ $i$ ]$entidade  $\leftarrow$  “adj”
  return taggedRel
```

Bloco de Código 4.8: Identificar descrições de condições clínicas

A variável “descricaoDeCondicoesClinicas” é uma lista composta pelas palavras da família das descrições das condições clínicas referidas acima.

<i>Input</i>			<i>Output</i>			
palavra	tag	lema	palavra	tag	lema	entidade
Visualiza	VMIP3S0	visualizar	Visualiza	VMIP3S0	visualizar	
se	PP3CN000	se	se	PP3CN000	se	
em	SPS00	em	em	SPS00	em	
o	DA0MS0	o	o	DA0MS0	o	
lobo	NCMS000	lobo	lobo	NCMS000	lobo	
esquerdo	AQ0MS0	esquerdo	esquerdo	AQ0MS0	esquerdo	
pequeno	AQ0MS0	pequeno	pequeno	AQ0MS0	pequeno	
quisto	AQ0MS0	quisto	quisto	AQ0MS0	quisto	
colóide	NCMS000	colóide	colóide	NCMS000	colóide	adj
com	SPS00	com	com	SPS00	com	
4mm	NCMS000	<unknow>	4mm	NCMS000	<unknow>	
.	Fp	.	.	Fp	.	

Tabela 4.8: Função IDENTIFICARDESCRICOESDECONDICOESCLINICAS *input/output*

4.5.4 Exceções

Existem ainda outras palavras da família “quisto/cisto” que podem pertencer tanto a condições clínicas como a uma descrição de uma condição clínica. Quando temos a seguinte frase, “Na hemitiróide direita identifica-se uma formação cística com 15mm, estável em relação ao exame anterior.”, “cística” é referente a uma condição clínica. Já quando temos a frase “No lobo esquerdo as duas maiores nodularidades têm características císticas, de provável natureza colóide e medem 3mm.”, “císticas” é referente a uma descrição de condição clínica. Para desfazer ambiguidades, em relação às palavras da família de “quístico”/“cístico” é necessário entender a qual das categorias pode pertencer cada uma das palavras, condição clínica ou descrição de condição clínica, dependendo do contexto. Para isso é preciso entender quais as frases que contêm palavras da família “nódulo” e verificar se à frente da palavra existem palavras da família “quístico”/“cístico”. Se existirem, “quístico”/“cístico” refere-se a uma descrição de uma condição clínica, se não, refere-se a uma condição clínica.

Para identificar as palavras da família “quístico”/“cístico” identificou-se, em todos os relatórios, as palavras que contêm “quisti”, “cisti”, “quísti” e “císti” (ver Anexo A.10).

```

function IDENTIFICAREXCECOES(taggedRel)
   $i \leftarrow 1$ 
  while  $i \leq \text{LENGTH}(\text{taggedRel})$  do
    if  $\text{taggedRel}[i].\text{palavra} \in \text{nodulos}$  then
      while  $\text{taggedRel}[i].\text{entidade} \neq \text{"end"}$  do
        if  $\text{taggedRel}[i].\text{palavra} \in \text{execoos}$  then
           $\text{taggedRel}[i].\text{entidade} \leftarrow \text{"adj"}$ 
           $i++$  break
        if  $\text{taggedRel}[i].\text{palavra} \in \text{execoos}$  then
           $\text{taggedRel}[i].\text{entidade} \leftarrow \text{"nome"}$ 
           $i++$ 
  return  $\text{taggedRel}$ 

```

Bloco de Código 4.9: Identificar exceções

A variável “nodulos” é uma lista composta pelas palavras da família “nódulo”, referidas acima.

A variável “execoos” é uma lista composta pelas palavras da família “quístico”/“cístico”, referidas acima.

<i>Input</i>			<i>Output</i>			
palavra	tag	lema	palavra	tag	lema	entidade
Identifica	VMIP3S0	identificar	Identifica	VMIP3S0	identificar	
se	PP3CN000	se	se	PP3CN000	se	
um	DI0MS0	um	um	DI0MS0	um	
nódulo	NCMS000	nódulo	nódulo	NCMS000	nódulo	
quístico	AQ0MS0	quístico	quístico	AQ0MS0	quístico	adj
com	SPS00	com	com	SPS00	com	
3mm	NCMS000	<unknow>	3mm	NCMS000	<unknow>	
em	SPS00	em	em	SPS00	em	
o	DA0MS0	o	o	DA0MS0	o	
lobo	NCMS000	lobo	lobo	NCMS000	lobo	
esquerdo	AQ0MS0	esquerdo	esquerdo	AQ0MS0	esquerdo	
.	Fp	.	.	Fp	.	

Tabela 4.9: Função IDENTIFICAREXCECOES *input/output*

4.6 Sequências

Depois de identificar cada uma das entidades, deve obter-se uma sequência que reflete a ordem de aparição das diferentes entidades juntamente com verbos, fim de frase e ocorrências relevantes.

Para isso, começamos por identificar as entidades, os verbos, o fim de frase e as ocorrências relevantes para posteriormente obter a sua ordem de aparição.

```

function IDENTIFICARENTIDADES(frasesRelevantes)
  taggedFrasesRelevantes  $\leftarrow$  {}
  for all  $fr \in$  frasesRelevantes do
    taggedRel  $\leftarrow$  TREETAGRELATORIOS( $fr$ )
    taggedRel$entidade  $\leftarrow$  ""
    taggedRel  $\leftarrow$  IDENTIFICARCOMPLEMENTOSDELUGAR(taggedRel)
    taggedRel  $\leftarrow$  IDENTIFICARVERBOS(taggedRel)
    taggedRel  $\leftarrow$  IDENTIFICARCONDICOESCLINICAS(taggedRel)
    taggedRel  $\leftarrow$  IDENTIFICARDESCRICAODECONDICOESCLINICAS(taggedRel)
    taggedRel  $\leftarrow$  IDENTIFICAREXCECOES(taggedRel)
    taggedRel  $\leftarrow$  IDENTIFICAROCORRENCIASRELEVANTES(taggedRel,  $fr$ )
    taggedRel  $\leftarrow$  IDENTIFICARFIMDEFRASE(taggedRel)
    taggedRel  $\leftarrow$  NUMERARENTIDADES(taggedRel)
    taggedFrasesRelevantes  $\leftarrow$  taggedFrasesRelevantes  $\cup$  taggedRel
  return taggedFrasesRelevantes

```

Bloco de Código 4.10: Identificação de entidades

“taggedRel\$entidade \leftarrow ”” adiciona o atributo “entidade” à variável “taggedRel”.

<i>Input</i>	<i>Output</i>			
	palavra	tag	lema	entidade
"Visualiza-se no lobo esquerdo pequeno quisto colóide com 4mm."	Visualiza	VMIP3S0	visualizar	verbo1
	se	PP3CN000	se	
	em	SPS00	em	c_lugar1
	o	DA0MS0	o	c_lugar1
	lobo	NCMS000	lobo	c_lugar1
	esquerdo	AQ0MS0	esquerdo	c_lugar1
	pequeno	AQ0MS0	pequeno	
	quisto	AQ0MS0	quisto	nome1
	colóide	NCMS000	colóide	adj1
	com	SPS00	com	
	4mm	NCMS000	<unknow>	num1
	.	Fp	.	end

Tabela 4.10: Função IDENTIFICARENTIDADES *input/output*

4.6.1 Ocorrências relevantes

Para identificar ocorrências relevantes basta obter uma lista destas e identificá-las.

```
function IDENTIFICAROCORRENCIASRELEVANTES(taggedRel, fraseRelevante)
  ocorrenciasRelevantes ← OCORRENCIASRELEVANTES(fraseRelevante)
  for i = 1 to LENGTH(taggedRel) do
    if taggedRel$.palavra[i] ∈ ocorrenciasRelevantes then
      taggedRel[i].entidade ← "num"
  return taggedRel
```

Bloco de Código 4.11: Identificar ocorrências relevantes

4.6.2 Fim de frase

De forma a identificar o final de uma frase, já que existem frases relevantes constituídas por duas frases, foi necessário identificar o fim destas. Para isso criamos uma entidade que identifique o fim de frase, ou seja, quando no atributo “palavra” existir um “.”, “!” ou “?” o atributo “entidade” deve ser “end”.

```
function IDENTIFICARFIMDEFRASE(taggedRel)
  for i = 1 to LENGTH(taggedRel) do
    if TEM(".", "!", "?", taggedRel[i].palavra) then
      taggedRel[i].entidade ← "end"
  return taggedRel
```

Bloco de Código 4.12: Identificar fim de frase

4.6.3 Verbos

O constituinte que designa o ser ou o objeto sobre o qual diretamente recai a ação expressa pelo verbo tem o nome de complemento direto. Este surge como a resposta à pergunta “O que é que?” [45]. Nos relatórios da tiroide, o complemento direto geralmente refere-se a uma condição clínica. O verbo, geralmente, antecede o complemento direto, tornando-se, assim, uma importante classe gramatical para a extração automática de condições clínicas.

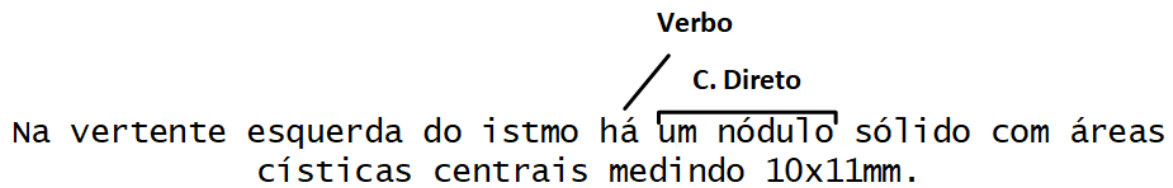


Figura 4.2: Exemplo complemento direto/condição clínica

Avaliando a ordem de aparição das entidades nas frases relevantes de 100 relatórios, e recorrendo a uma implementação do algoritmo *GSP* do *Weka* [47], como podemos ver na imagem 4.3, em 217 ocorrências de condições clínicas, a sequência verbo condição clínica (“nome”) acontece 187 vezes. O verbo torna-se, assim, uma importante classe gramatical para a extração automática de condições clínicas. Para identificar um verbo só é preciso selecionar todas as palavras classificadas pelo *POSTagging* como verbos.

```
- 2-sequences
```

[1]	<{nome}{verbo}>	(176)
[2]	<{verbo}{nome}>	(187)

Figura 4.3: Numero de ocorrências *verbo* → condição clínica ("nome")

```
function IDENTIFICARVERBOS(taggedRel)
  for i = 1 to LENGTH(taggedRel) do
    if CHARAT(taggedRel[i]$tag, 1) == "V" then
      taggedRel[i]$categoria ← "verbo"
  return taggedRel
```

Bloco de Código 4.13: Identificação de verbos

A função “CHARAT” retorna um carácter que se encontra numa posição dada de uma *string*. O primeiro parâmetro é a *string*, o segundo a posição do carácter que se pretende obter.

4.6.4 Ordem de ocorrência

As entidades podem aparecer várias vezes em cada uma das frases relevantes, por isso torna-se imperativo numerar estas de forma a perceber a sua ordem de aparição. Assim, é possível obter uma verdadeira sequência que retrata a ordem de aparição das entidades.

```

function NUMERARENTIDADES(taggedRel)
  cLugar ← 1
  verbo ← 1
  nome ← 1
  adj ← 1
  num ← 1
  i ← 1
  while i ≤ LENGTH(taggedRel) do
    if taggedRel[i].Sentidade == "c_lugar" then
      while taggedRel[i].Sentidade == "c_lugar" do
        taggedRel[i].Sentidade ← CONCAT("c_lugar", cLugar)
        i ++
        cLugar ++
      if taggedRel[i].Sentidade == "verbo" then
        taggedRel[i].Sentidade ← CONCAT("verbo", verbo)
        verbo ++
      else if taggedRel[i].Sentidade == "nome" then
        taggedRel[i].Sentidade ← CONCAT("nome", nome)
        nome ++
      else if taggedRel[i].Sentidade == "adj" then
        taggedRel[i].Sentidade ← CONCAT("adj", adj)
        adj ++
      else if taggedRel[i].Sentidade == "num" then
        taggedRel[i].Sentidade ← CONCAT("num", num)
        num ++
      i ++
  return taggedRel

```

Bloco de Código 4.14: Numerar entidades

<i>Input</i>				<i>Output</i>			
palavra	tag	lema	entidade	palavra	tag	lema	entidade
Visualiza	VMIP3S0	visualizar	verbo	Visualiza	VMIP3S0	visualizar	verbo1
se	PP3CN0	se		se	PP3CN0	se	
em	SPS0	em	c_lugar	em	SPS0	em	c_lugar1
o	DA0MS0	o	c_lugar	o	DA0MS0	o	c_lugar1
lobo	NCMS00	lobo	c_lugar	lobo	NCMS0	lobo	c_lugar1
esquerdo	AQ0MS0	esquerdo	c_lugar	esquerdo	AQ0MS0	esquerdo	c_lugar1
pequeno	AQ0MS0	pequeno		pequeno	AQ0MS0	pequeno	
quisto	AQ0MS0	quisto	nome	quisto	AQ0MS0	quisto	nome1
colóide	NCMS0	colóide	adj	colóide	NCMS0	colóide	adj1
com	SPS00	com		com	SPS00	com	
4mm	NCMS00	<uk>	num	4mm	NCMS00	<uk>	num1
.	Fp	.	end	.	Fp	.	end

Tabela 4.11: Função NUMERARENTIDADES *input/output*

4.6.5 Obter sequência

Para obter uma sequência extraímos as entidades por ordem de ocorrência. A frase da Tabela 4.11 “Visualiza-se no lobo esquerdo pequeno quisto colóide com 4mm”, deve produzir a seguinte sequência: “verbo1, c_lugar1, nome1, adj1, num1, end”. Sendo, “c_lugar” o complemento circunstancial de lugar, o “adj” a descrição da condição clínica, o “nome” uma condição clínica, “verbo” um verbo, “end” o fim de uma frase e “num” uma ocorrência relevante.

```

function OBTERSEQUENCIA(taggedRel)
  sequencia ← {}
  i ← 1
  while i ≤ (LENGTH(taggedRel) – 1) do
    if taggedRel[i]$entidade == “c_lugar” then
      sequencia ← sequencia ∪ taggedRel[i]$entidade
      while taggedRel[i]$entidade == “c_lugar” do
        i ++
    if taggedRel[i]$entidade ≠ “” then
      sequencia ← sequencia ∪ taggedRel[i]$entidade
      i ++
  return sequencia

```

Bloco de Código 4.15: Obter sequência

<i>Input</i>				<i>Output</i>
palavra	tag	lema	entidade	
Visualiza	VMIP3S0	visualizar	verbo1	"verbo1, c_lugar1, nome1, adj1, num1"
se	PP3CN000	se		
em	SPS00	em	c_lugar1	
o	DA0MS0	o	c_lugar1	
lobo	NCMS000	lobo	c_lugar1	
esquerdo	AQ0MS0	esquerdo	c_lugar1	
pequeno	AQ0MS0	pequeno		
quisto	AQ0MS0	quisto	nome1	
colóide	NCMS000	colóide	adj1	
com	SPS00	com		
4mm	NCMS000	<unknow>	num1	
.	Fp	.	end	

Tabela 4.12: Função OBTERSEQUENCIA *input/output*

Posteriormente, obtida a sequência, identificamos quais as entidades associadas a uma ocorrência relevante.

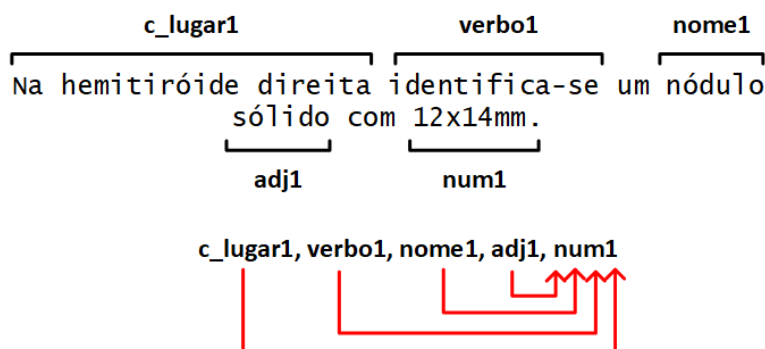


Figura 4.4: Entidades associadas a uma ocorrência relevante

Por exemplo, na frase da Figura 4.4 “Na hemitiróide direita identifica-se um nódulo sólido com 12x14mm.” sabemos que à ocorrência relevante “num1” estão associadas as entidades “c_lugar”, “nome1” e “adj”. Estes devem ser as entidades a serem extraídas automaticamente, ou seja, devem ser extraídas automaticamente as entidades associadas às ocorrências relevantes. Desta forma, é possível extrair, por completo, os termos relevantes.

A sequência a ser analisada para extração automática das entidades é a sequência que se encontra depois do primeiro “end”, ou seja, a sequência da última frase, onde ocorre a ocorrência relevante. A sequência completa só é usada se a entidade condição clínica não existir, na última frase, à esquerda da ocorrência relevante. Em ambos os casos, no final da última frase é identificada a entidade “end”. Esta é eliminada, pois não se demonstrou relevante para a extração de entidades associadas a ocorrências relevantes, como demonstrado no Bloco de Código 4.15.


```

function ULTIMASEQUENCIA(sequencia)
  for  $i = 1$  to LENGTH(sequencia) do
    if sequencia[i] == "end" then
      return sequencia[(i + 1) : LENGTH(sequencia)]

```

Bloco de Código 4.16: Obter última sequência

<i>Input</i>	<i>Output</i>
"verbo1, nome1, num1, end, c_lugar1, verbo2, nome2, adj1, adj2, verbo2, num2"	"c_lugar1, verbo2, nome2, adj1, adj2, verbo3, num2"

Tabela 4.13: Função ULTIMASEQUENCIA *input/output*

4.7 Extração automática das entidades associadas às ocorrências relevantes

Analisando 170 sequências de frases relevantes, correspondentes a 100 relatórios da tiroide e a 266 ocorrências relevantes, foi possível obter manualmente as sequências relevantes. Uma sequência relevante é uma sequência onde consta entidades associadas a uma ocorrência relevante. Por exemplo, para a ocorrência relevante “num1” da Figura 4.5 obtemos a seguinte sequência relevante: “c_lugar1, verbo2, nome1, adj1, adj2”.

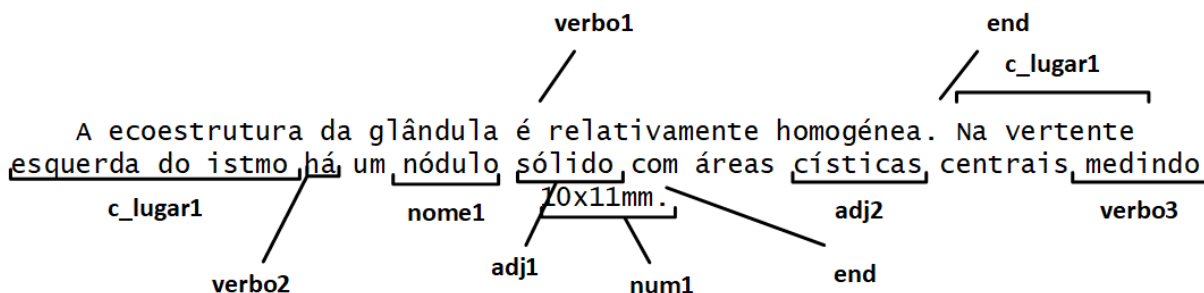


Figura 4.5: Entidades associadas a uma ocorrência relevante

Depois de analisadas as 266 sequências relevantes chegamos à conclusão que para extrair automaticamente as entidades associadas a uma ocorrência relevante é necessário tratar, de forma independente, cada uma das entidades (condição clínica, descrição de condição clínica e complemento circunstancial de lugar).

4.7.1 Extração de condições clínicas

Uma ocorrência relevante tem, obrigatoriamente, uma condição clínica. Estas apresentam-se à esquerda da ocorrência relevante, na maior parte dos casos, e antecedidas de um verbo, ou seja, fazem parte do complemento direto. O complemento direto poderia ser importante para identificar situações ambíguas no reconhecimento de entidades, nomeadamente na distinção entre descrições de condições clínicas e condições clínicas. Por exemplo, na frase “Identifica-se uma formação cística com 2mm” “cística” é referente a uma condição clínica, já quando temos a frase “No lobo esquerdo, identificam-se nodularidades com características císticas medindo 5mm.”, “císticas” é referente à descrição de condição clínica. Tanto na primeira frase como na segunda, o verbo antecede a condição clínica, ou seja, a condição clínica pertence ao complemento direto, embora isto não aconteça para todos os casos. Por exemplo, a frase “Conclusão dois nódulos na hemitiroide direita respetivamente com 6mm e 11mm.” a condição clínica “nódulos” não precede de um verbo, nem faz parte do complemento direto. De forma a evitar más interpretações, foi extraída a condição clínica que se encontra mais próxima e à esquerda da ocorrência relevante, como é exemplificado na Figura 4.6. Sendo que, para distinguir condições clínicas e descrições de condições clínicas foi utilizado o método mencionado em 4.5.4.

c_lugar1, verbo1, nome1, adj1, adj2, num1



 Um diagrama que mostra a extração automática de uma condição clínica a partir de uma frase. A frase é representada por uma linha horizontal com o texto "c_lugar1, verbo1, nome1, adj1, adj2, num1". Abaixo do texto, há uma linha vermelha que se estende desde o início da frase até o final da palavra "num1", terminando com uma seta vermelha apontando para cima.

Figura 4.6: Extração automática de condição clínica, uma frase

Existem casos onde a metodologia para a extração automática da condição clínica falha, ou seja, quando na frase da ocorrência relevante, à esquerda da ocorrência não existe uma condição clínica. Se numa primeira tentativa de extração não conseguirmos extrair uma condição clínica, então temos de verificar se na frase anterior, com a mesma metodologia (à esquerda e a mais próxima da ocorrência relevante), conseguimos extrair uma condição clínica, como é exemplificado na Figura 4.7.

verbo3, nome1, end, c_lugar1, adj1, num1, adj2



 Um diagrama que mostra a extração automática de uma condição clínica a partir de duas frases. A frase é representada por uma linha horizontal com o texto "verbo3, nome1, end, c_lugar1, adj1, num1, adj2". Abaixo do texto, há uma linha vermelha que se estende desde o início da frase até o final da palavra "num1", terminando com uma seta vermelha apontando para cima.

Figura 4.7: Extração automática de condição clínica, duas frases

Se mesmo assim não conseguirmos extrair uma condição clínica, então a ocorrência relevante não estará associada a nenhuma condição clínica e, portanto, não estará associada a nenhuma entidade. Esta ocorrência relevante não pertencerá, muito provavelmente, a nenhum termo relevante. É o exemplo das frases "Ecografia da tiróide técnica foram realizados ecotomogramas sagitais e transversais da glândula tiróide. Relatório a tiróide localiza-se na sua topografia habitual e tem dimensões conservadas (lobo direito 56x17x12mm; lobo esquerdo 57x14x13mm, de diâmetros longitudinal, antero-posterior e transversal; istmo 3mm)." de onde conseguimos

extrair a sequência “verbo2, end, num1”. A ocorrência relevante “num1”, do exemplo anterior, não apresenta nenhuma condição clínica associada.

```

function EXTRAIRCONDICAOCLINICA(ORrelevante,sequencia)
   $i \leftarrow 1$ 
  while  $i = 1 \leq \text{LENGTH}(sequencia)$  do
    if  $sequencia[i] == ORrelevante$  then
       $i - -$ 
      while  $i > 0$  do
        if  $\text{TEM}(\text{"nome}[0 - 9]", sequencia[i])$  then
          return  $sequencia[i]$ 
         $i - -$ 
      return  $\{\}$ 
     $i + +$ 

```

Bloco de Código 4.17: Extrair condição clínica

<i>Input</i>	<i>Output</i>
"num1"; "c_lugar1, verbo1, nome1, adj1, adj2, verbo2, num1"	"nome1"

Tabela 4.14: Função EXTRAIRCONDICAOCLINICA *input/output*

4.7.2 Extração de descrições de condições clínicas

Para fazer a extração das descrições das condições clínicas, é essencial saber qual a posição na sequência da ocorrência relevante, bem como da condição clínica associada à ocorrência relevante. As descrições das condições clínicas variam consoante as posições destas na sequência. Para entender qual a posição das condições clínicas, descrições das condições clínicas e ocorrência relevante, removemos das sequências todos os outros elementos (verbos e complementos circunstanciais de lugar).

```

function ELIMINARVERBOSDECLUGAR(sequencia)
   $novaSequencia \leftarrow \{\}$ 
  for all  $s \in sequencia$  do
    if  $\neg \text{TEM}(\text{"verbo}[0 - 9]", s)$  or  $\neg \text{TEM}(\text{"c_lugar}[0 - 9]", s)$  then
       $novaSequencia \leftarrow novaSequencia \cup s$ 
  return  $novaSequencia$ 

```

Bloco de Código 4.18: Eliminar verbos e complementos circunstanciais de lugar

<i>Input</i>	<i>Output</i>
"c_lugar1, verbo1, nome1, adj1, num1"	"nome1, adj1, num1"

Tabela 4.15: Função ELEMENARVERBOSECLUGAR *input/output*

Sabendo de antemão qual é a condição clínica associada a uma ocorrência relevante é possível extrair com mais precisão as descrições das ocorrências relevantes, na sequência, que normalmente surgem entre a condição clínica e a ocorrência relevante. Para verificar, é preciso entender qual é a primeira descrição de condição clínica que surge à esquerda da ocorrência relevante. Depois de encontrada, são extraídas todas as descrições à esquerda, até aparecer um elemento diferente de descrição (condição clínica ou uma ocorrência relevante), como está exemplificado para a ocorrência relevante “num1” na Figura 4.8 e para a ocorrência relevante “num2” da Figura 4.9.

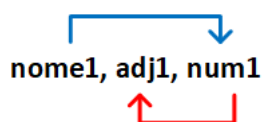


Figura 4.8: Extração automática de descrições clínicas, exemplo 1

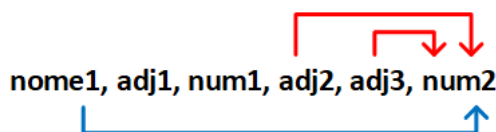


Figura 4.9: Extração automática de descrições clínicas, exemplo 2

Existem exceções, nomeadamente quando a descrição da condição clínica não se encontra entre a condição clínica e a ocorrência relevante. Nestes casos, muito provavelmente, a descrição da condição clínica encontra-se no fim da sequência, ou seja, é necessário saber qual a primeira descrição de condição clínica que se encontra à direita da ocorrência relevante e verificar se é precedida, apenas e só, de descrições de condições clínicas (ex. Figura 4.10). Não sendo precedida apenas de descrições de condições clínicas, então podemos dizer que não existe uma descrição de condição clínica associada à ocorrência relevante, sendo impossível extraí-la, como é o exemplo da ocorrência relevante “num1” na sequência “nome1, num1, adj2, num2”.

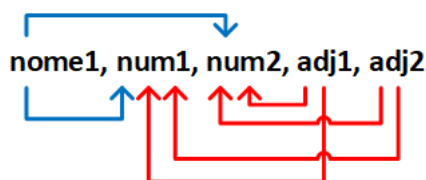


Figura 4.10: Extração automática de descrições clínicas, exemplo 3

Por fim, se uma ocorrência relevante for a última ocorrência da sequência e à direita desta só existirem descrições de condições clínicas, então estas estão associadas à última ocorrência

relevante da sequência, como está exemplificado para a ocorrência relevante “num2” na Figura 4.11.

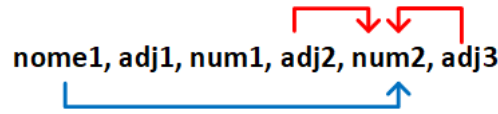


Figura 4.11: Extração automática de descrições clínicas, exemplo 4

```

function EXTRACAODeDESCRICOESDeCONDICOESCLINICAS(ORrelevante, CClinica, sequen-
cia)
    descricaoDeCondicaoClinica ← {}
    sequencia ← ELEMENARVERBOSECLUGAR(sequencia)
    for i = 1 to LENGTH(sequencia) do
        if sequencia[i] == ORrelevante then
            j ← i
            while sequencia[j] ≠ CClinica do
                if TEM("adj[0 – 9]", sequencia[j]) then
                    while TEM("adj[0 – 9]", sequencia[j]) do
                        descricaoDeCondicaoClinica ← descricaoDeCondicaoClinica ∪
sequencia[j]
                        j – –
                    break
                j – –
            while i < LENGTH(sequencia) do
                if TEM("adj[0 – 9]", sequencia[i]) then
                    inicio ← i
                    while TEM("adj[0 – 9]", sequencia[i]) do
                        i + +
                    if i == LENGTH(sequencia) then
                        descricaoDeCondicaoClinica ← descricaoDeCondicaoClinica ∪
sequencia[inicio : i]
                    break
                i + +
            return descricaoDeCondicaoClinica
i + +

```

Bloco de Código 4.19: Extração de descrições de condições clínicas

<i>Input</i>	<i>Output</i>
"num1"; "nome1"; "c_lugar1, verbo1, nome1, adj1, adj2, verbo2, num1"	"adj1, adj2"

Tabela 4.16: Função EXTRACAODEDESCRICAODECONDICOESCLINICAS *input/output*

4.7.3 Extração de complementos circunstanciais de lugar

Dada a volatilidade do complemento circunstancial de lugar, ou seja, este pode aparecer em qualquer parte da frase, a extração deste torna-se mais difícil. Para extrair o complemento circunstancial de lugar removemos das sequências: verbos, condições clínicas e descrições de condições clínicas.

```

function ELIMINARVERBOSECCLINICASEDCCLINICAS(sequencia)
  novaSequencia  $\leftarrow$  {}
  for all  $s \in$  sequencia do
    if  $\neg$ TEM("verbo[0 – 9]", s) or  $\neg$ TEM("nome[0 – 9]", s) or  $\neg$ TEM("adj[0 – 9]") then
      novaSequencia  $\leftarrow$  novaSequencia  $\cup$  s
  return novaSequencia

```

Bloco de Código 4.20: Eliminar verbos, condições clínicas e descrição de condições clínicas

<i>Input</i>	<i>Output</i>
"c_lugar1, verbo1, nome1, adj1, num1"	"c_lugar1, nome1"

Tabela 4.17: Função ELEMENARVERBOSECCLINICASEDCCLINICAS *input/output*

Analisando as sequências com complementos circunstanciais de lugar e ocorrências relevantes observa-se que os complementos de lugar ocorrem, geralmente, à direita da ocorrência relevante se a sequência terminar em complemento de lugar ou à esquerda se a sequência não terminar em complemento de lugar, como é exemplificado nas Figuras 4.12, 4.13 e 4.14.

num1, c_lugar1, num2, c_lugar2

Figura 4.12: Extração automática de complementos circunstanciais de lugar, exemplo 1

c_lugar1, num1, c_lugar2, num2

Figura 4.13: Extração automática de complementos circunstanciais de lugar, exemplo 2

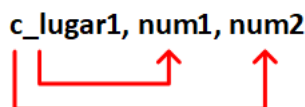


Figura 4.14: Extração automática de complementos circunstanciais de lugar, exemplo 3

Existem ainda sequências onde é impossível encontrar um complemento de lugar à esquerda da ocorrência relevante, pois esta é a primeira da sequência. Para esta ocorrência relevante, o complemento de lugar associado passa a ser o primeiro que aparecer à direita, como é exemplificado na Figura 4.15.

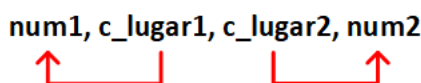


Figura 4.15: Extração automática de complementos circunstanciais de lugar, exemplo 4

Para uma ocorrência relevante é apenas extraído um complemento de lugar, embora existam casos onde estas têm mais de um. Nestes casos, normalmente, os complementos de lugar são referentes à mesma zona da tiroide, sendo que uma é referente a uma zona mais abrangente da tiroide, como é demonstrado na Figura 4.16.

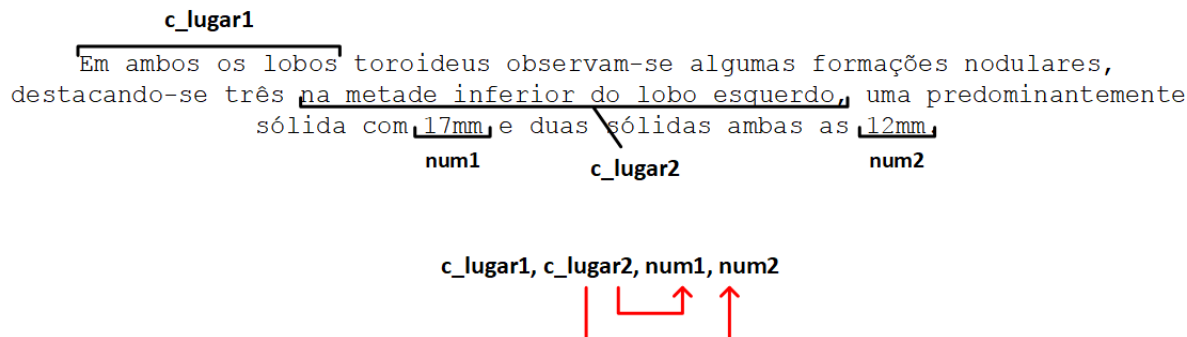


Figura 4.16: Extração automática de complementos circunstanciais de lugar, exemplo 5

```

function EXTRACAOEDECCDELUGAR(ORrelevante, sequencia)
    sequencia ← ELIMINARVERBOSECCCLINICASECCCLINICAS(sequencia)
    if TEM("c_lugar[0 – 9]", sequencia[LENGTH(sequencia)]) then
        direita ← true
    else
        direita ← false
    for i = 1 to LENGTH(sequencia) do
        if sequencia[i] == ORrelevante then
            if direita or i == 1 then
                while ¬TEM("c_lugar[0 – 9]", sequencia[i]) do
                    i ++
                return sequencia[i]
            else
                while ¬TEM("c_lugar[0 – 9]", sequencia[i]) do
                    i --
                return sequencia[i]
    return {}

```

Bloco de Código 4.21: Extração de complementos circunstanciais de lugar

<i>Input</i>	<i>Output</i>
"num1"; "c_lugar1, verbo1, nome1, adj1, adj2, verbo2, num1"	"c_lugar1"

Tabela 4.18: Função EXTRACAOEDECCDELUGAR *input/output*

4.8 Extração de termos relevantes

Por fim, resta extrair os termos relevantes, ou seja, ocorrências relevantes e entidades associadas a estas. Para isso, começamos por tratar o texto, selecionar as frases relevantes e fazer o *TreeTagg* das frases relevantes. De seguida, para todas as frases relevantes obtemos a sequência, a sequência da última frase e as ocorrências relevantes da última frase, para cada uma dessas ocorrências é extraído um termo relevante. Numa primeira fase, para cada uma das ocorrências relevantes tentamos extrair a condição clínica, na última frase. Se esta existir, são extraídas as outras entidades na última frase, se não, são extraídas as entidades na sequência, ou seja, na última frase e na frase anterior. Por fim, identificamos a(s) palavra(s) correspondente(s) à entidade extraída na sequência.


```

function EXTRAIRTERMOSRELEVANTES(taggedFrasesRelevantes)
  for all tfr  $\in$  taggedFrasesRelevantes do
    sequencia  $\leftarrow$  OBTERSEQUENCIA(tfr)
    ultimaSequencia  $\leftarrow$  ULTIMASEQUENCIA(sequencia)
    oRelevantes  $\leftarrow$  ENTORELEVANTE(ultimaSequencia)
    termosRelevantes  $\leftarrow$  {}
    for all or  $\in$  oRelevante do
      termo$cClinica  $\leftarrow$  EXTRAIRCONDICAOCLINICA(or, ultimaSequencia)
      if termo$cClinica == {} then
        termo$cClinica  $\leftarrow$  EXTRAIRCONDICAOCLINICA(or, sequencia)
        if termo$cClinica == {} then
          termosRelevantes  $\leftarrow$  termosRelevantes  $\cup$  {}
        else
          termo$dCClinica  $\leftarrow$ 
EXTRACAODEDESCRICAODECONDICAOCLINICA(or, termo$cClinica, sequencia)
          termo$cLugar  $\leftarrow$  EXTRACAODECCDELUGAR(or, sequencia)
        else
          termo$dCClinica  $\leftarrow$ 
EXTRACAODEDESCRICAODECONDICAOCLINICA(or, termo$cClinica, ultimaSequencia)
          termo$cLugar  $\leftarrow$  EXTRACAODECCDELUGAR(or, ultimaSequencia)
          termoEmTexto$cClinica  $\leftarrow$  OBTERTEXTO(tfr, termo$cClinica)
          termoEmTexto$dCClinica  $\leftarrow$  OBTERTEXTO(tfr, termo$dCClinica)
          termoEmTexto$cLugar  $\leftarrow$  OBTERTEXTO(tfr, termo$cLugar)
          termoEmTexto$or  $\leftarrow$  OBTERTEXTO(trf, or)
          termosRelevantes  $\leftarrow$  termosRelevantes  $\cup$  termoEmTexto
    return termosRelevantes

```

Bloco de Código 4.22: Extrair termos relevantes

```

function ENTORELEVANTE(sequencia)
  ORelevante  $\leftarrow$  {}
  for all s  $\in$  sequencia do
    if TEM("num", s) then
      ORelevante  $\leftarrow$  ORelevante  $\cup$  s
  return ORelevante

```

Bloco de Código 4.23: Extrair ocorrência relevante de sequência

```

function OBTERTEXTO(taggedRel, entidade)
  palavras  $\leftarrow$  {}
  for  $i = 1$  to LENGTH(taggedRel) do
    if taggedRel[ $i$ ]$entidade == entidade then
      palavras  $\leftarrow$  palavras  $\cup$  taggedRel[ $i$ ]$palavra
  return palavras

```

Bloco de Código 4.24: Obter texto

<i>Input</i>				<i>Output</i>			
palavra	tag	lema	entidade	cClinica	dCClinica	cLugar	OR
Visualiza	VMIP3S0	visualizar	verbo	"quisto"	"colóide"	"em o lobo esquerdo"	"4mm"
se	PP3CN000	se					
em	SPS00	em	c_lugar				
o	DA0MS0	o	c_lugar				
lobo	NCMS000	lobo	c_lugar				
esquerdo	AQ0MS0	esquerdo	c_lugar				
pequeno	AQ0MS0	pequeno					
quisto	AQ0MS0	quisto	nome				
colóide	NCMS000	colóide	adj				
com	SPS00	com					
4mm	NCMS000	<unknow>	num				
.	Fp	.					

Tabela 4.19: Função EXTRAIRTERMOSRELEVANTES *input/output*

Capítulo 5

Avaliação

Neste capítulo é apresentada a avaliação do sistema desenvolvido.

Para avaliar a eficácia do sistema é necessário avaliar a eficácia da extração automática. Para isso, avaliamos a extração automática de cada uma das entidades comparando os resultados da extração automática com relatórios previamente anotados, sendo que para cada entidade os métodos de avaliação podem ser diferentes. No final, para avaliar o sistema identificamos em quantas ocorrências relevantes o sistema extraiu, erradamente, os termos relevantes.

5.1 Conjunto de dados

O primeiro conjunto de dados é constituído por 100 relatórios escolhidos aleatoriamente dos 1299 relatórios que contêm uma ou mais ocorrências relevantes. Este conjunto foi referido acima e foi utilizado para obter as sequências relevantes, que, por sua vez, geraram o conjunto de regras utilizadas no sistema para a extração automática das diferentes entidades.

O segundo conjunto também é constituído por 100 relatórios escolhidos aleatoriamente dos 1299 que contêm ocorrências relevantes e serve para validar o conjunto de regras geradas pelo primeiro conjunto.

O primeiro conjunto de dados corresponde a um total de 170 frases relevantes e a 266 ocorrências relevantes. O segundo conjunto corresponde a um conjunto de 147 frases relevantes e a 257 ocorrências relevantes. Em ambos os conjuntos, a uma ocorrência relevante está associada uma sequência e uma sequência relevante.

Para a frase “Identificam-se dois pequenos nódulos, um no pólo inferior do lobo esquerdo, sólido e hipocogénico, medindo 3mm, e outro no pólo inferior do lobo direito, quístico, medindo 4mm.” identificamos as entidades representadas na Figura 5.1 e para as frases “Continuam a identificar-se algumas formações nodulares de natureza mista, bilateralmente. Destacam-se as de maiores dimensões no terço médio de ambos os lobos, isoecogénicas, predominantemente sólidas, medindo à direita 9mm e à esquerda 8mm.” identificamos as entidades representadas na Figura

5.2. Na Figura 5.1 é apresentado um exemplo em que a condição clínica associada às ocorrências relevantes se encontra na própria frase. Na Figura 5.2 é apresentado um exemplo em que a condição clínica associada às ocorrências relevantes se encontra na frase anterior.

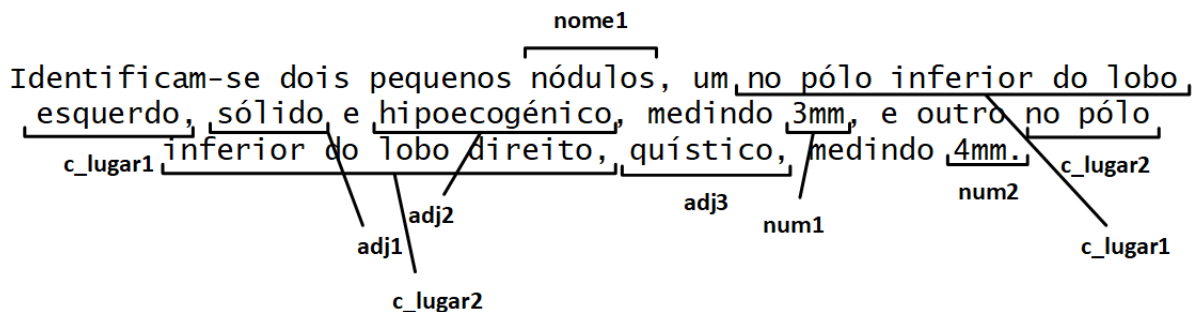


Figura 5.1: Identificação de entidades uma frase relevante

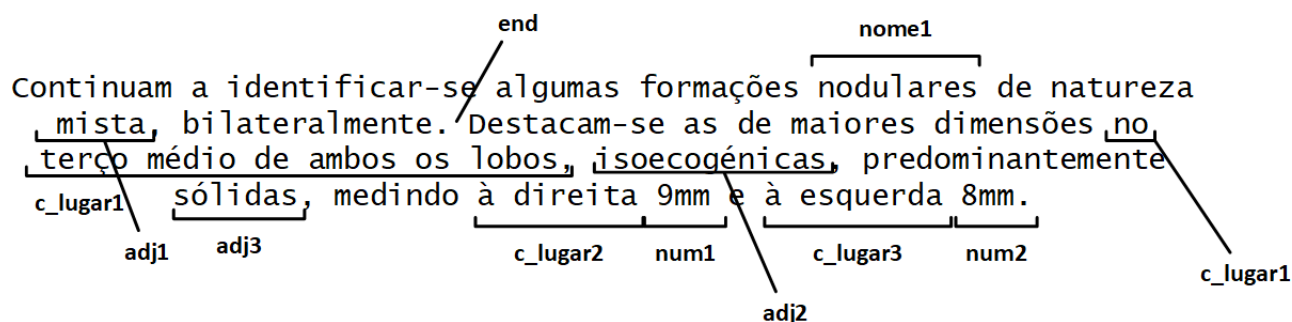


Figura 5.2: Identificação de entidades duas frases relevantes

Para avaliar a extração automática de termos para as frases das Figuras 5.1 e 5.2 obtemos para cada ocorrência relevante a sequência da(s) frase(s), as suas sequências relevantes e os resultados da extração automática de cada uma das entidades (condição clínica, descrição de condição clínica e complemento circunstancial de lugar).

Nas primeiras duas linhas da Tabela 5.1 temos as duas ocorrências relevantes do exemplo da Figura 5.1 e nas duas últimas linhas temos as duas ocorrências relevantes do exemplo da Figura 5.2.

OR		Sequências		Extração automática de termos		
		Sequência	Sequência Relevante	C. Clí-nica	D. C. Clínica	C. C. Lu-gar
3mm	num1	nome1, c_lugar1, adj1, adj2, num1, c_lugar2, adj3, num2	nome1, c_lugar1, adj1, adj2	nome1	adj1 adj2	c_lugar1
4mm	num2	nome1, c_lugar1, adj1, adj2, num1, c_lugar2, adj3, num2	nome1, c_lugar2, adj3	nome1	adj3	c_lugar2
9mm	num1	nome1, adj1, end, c_lugar1, adj2, adj3, c_lugar2, num1, c_lugar3, num2	nome1, adj1, adj2, adj3, c_lugar2	nome1	adj1 adj2 adj3	c_lugar2
8mm	num2	nome1, adj1, end, c_lugar1, adj2, adj3, c_lugar2, num1, c_lugar3, num2	nome1, adj1, adj2, adj3, c_lugar3	nome1	adj1 adj2 adj3	c_lugar3

Tabela 5.1: Conjunto de dados exemplificativos

Na primeira coluna, “OR” é constituído pela ocorrência relevante por extenso e pela identificação da ocorrência relevante. Por exemplo, a ocorrência relevante “3mm” da Figura 5.1 é identificada como sendo “num1”. Sendo apresentado para essa ocorrência a sequência da frase relevante “nome1, c_lugar1, adj1, adj2, num1, c_lugar2, adj3, num2”, a sequência relevante “nome1, c_lugar1, adj1, adj2” (“nome1” como “nódulos”, “c_lugar” como “no pólo inferior do lobo esquerdo”, “adj1” como “sólido” e “adj2” como “hipoecogénico”) e as entidades extraídas (“nome1”, “adj1”, “adj2” e “c_lugar1”).

Para avaliar o sistema verificamos se as entidades extraídas automaticamente estão contidas no conjunto da sequência relevante.

As sequências relevantes no primeiro conjunto foram anotadas manualmente, analisando a frase relevante ou as frases relevantes de cada ocorrência relevante. Já no segundo conjunto, as sequências relevantes foram geradas automaticamente pelo sistema e posteriormente corrigidas.

5.2 Medidas de avaliação

Para avaliar o sistema existem várias métricas: *Accuracy*, *Precision*(π), *Recall*(ρ), e *F-Measure* (F1).

De forma a avaliar o processo de extração de entidades, foram definidos os seguintes resultados: verdadeiro positivo (VP), falso positivo (FP), falso negativo (FN) e verdadeiro negativo (VN). VP para ocorrências relevantes que foram classificadas como contendo um termo e esse termo foi extraído pelo sistema; FP para ocorrências relevantes que foram classificadas como não contendo um termo e o sistema extraiu um termo; FN para ocorrências relevantes que foram classificadas como contendo um termo e o sistema não extraiu um termo e VN para ocorrências relevantes que foram classificadas como não contendo um termo e o sistema não extraiu um termo. Segue o exemplo na Tabela 5.2.

Resultados		Termo	
		Presente	Ausente
Extração Automática de Termo	Presente	VP	FP
	Ausente	FN	VN

Tabela 5.2: Resultados de Avaliação

A *Accuracy* foi usada para avaliar a extração automática das condições clínicas e dos complementos circunstanciais de lugar, pois para estas entidades a extração automática deve extrair para cada ocorrência relevante, como referido acima, uma ou nenhuma condição clínica ou um ou nenhum complemento circunstancial de lugar.

Accuracy é o número total de extrações corretas (VP+VN) sobre o número total de extrações (VP+FP+FN+VN).

$$Accuracy = \frac{VP+VN}{VP+FP+FN+VN} \quad (5.1)$$

Já para as descrições das condições clínicas é possível extrair zero ou várias descrições. Sendo assim, utilizamos *Precision*, *Recall*, e *F-Measure* (F1).

Precision é a cardinalidade da intersecção do conjunto extraído automaticamente (CAut) com o conjunto anotado (CAnt) ($\#(CAut \cap CAnt)$) sobre a cardinalidade do conjunto extraído automaticamente ($\#CAut$).

$$\pi = \frac{(\#(CAut \cap CAnt))}{\#CAut} \quad (5.2)$$

Recall é a cardinalidade da intersecção do conjunto extraído automaticamente (CAut) com o conjunto anotado (CAnt) $(\#(CAut \cap CAnt))$ sobre a cardinalidade do conjunto anotado $(\#CAnt)$.

$$\rho = \frac{(\#(CAut \cap CAnt))}{\#CAnt} \quad (5.3)$$

No *F-Measure* combinamos *Recall* e *Precision*, onde

$$F1 = \frac{2\rho\pi}{\rho+\pi} \quad (5.4)$$

5.3 Resultados e Análise

Para cada conjunto obtiveram-se os seguintes resultados:

5.3.1 Primeiro conjunto de dados

5.3.1.1 Extração de condições clínicas

Resultados		Termo	
		Presente	Ausente
Extração Automática de Termo	Presente	216	4
	Ausente	1	45

Tabela 5.3: Resultados da extração automática condições clínicas, primeiro conjunto

$$Accuracy = \frac{216+45}{216+4+1+45} = 0.981 \quad (5.5)$$

A extração automática de condições clínicas falhou em 5 das 266 ocorrências relevantes. A *accuracy* na extração automática de condições clínicas é de 0.981 o que significa que menos de 2% das condições clínicas extraídas foram erradamente extraídas.

A extração automática de condições clínicas nos casos identificados como FP falha quando é erradamente identificada uma ocorrência relevante. Neste casos, o sistema de extração procura uma condição clínica na frase anterior. Se a encontrar, vai extrair a condição clínica da frase anterior embora, a condição clínica da frase anterior não pertença à ocorrência identificada como relevante, aliás, nestes casos, a condição clínica associada, tal como as outras entidades associadas, não existem. Por exemplo, nas frases “Os nódulos acima referidos apresentam uma estrutura predominantemente quística. Os diâmetros da hemitiroide direita são: longitudinal 4,7cm, transversal 1,7cm, antero-posterior 1,3cm; diâmetros da hemitiroide esquerda: longitudinal 4,6cm, transversal 1, 9cm, antero-posterior 9mm.” a condição clínica extraída automaticamente para a ocorrência relevante “9mm” é identificada, erradamente, como sendo “nódulos” da frase anterior.

5.3.1.2 Extração de complementos circunstanciais de lugar

Resultados		Termo	
		Presente	Ausente
Extração Automática de Termo	Presente	190	1
	Ausente	1	74

Tabela 5.4: Resultados da extração automática de complementos circunstanciais de lugar, primeiro conjunto

$$Accuracy = \frac{190+74}{190+1+1+74} = 0.992 \quad (5.6)$$

A extração automática de complementos circunstanciais de lugar falhou em 2 das 266 ocorrências relevantes. A *accuracy* na extração automática de condições clínicas é de 0.992 o que significa que menos de 1% das condições clínicas extraídas foram erradamente extraídas, sendo o falso positivo a ocorrência relevante “6mm” da Figura 5.3 e o falso negativo a ocorrência relevante “9mm” da frase da Figura 5.4.

A sua ecoestrutura é heterogênea, alterada por múltiplas formações quísticas cujas dimensões não excedem os 6mm, salientando-se ainda nódulo no lobo direito, que mede 27mm, segundo informação da paciente já submetido a punção aspirativa, recomendando-se manutenção da avaliação ecográfica.

num1

c_lugar1

num2

Figura 5.3: Identificação de entidades frase 3

O istmo é proeminente, com um espessamento do istmo, de aproximadamente 5mm, definindo-se nesta região ístmica nódulo, com cerca de 9mm, e outros três com cerca de 4, e 5mm, no lobo esquerdo, aconselhando-se manutenção da vigilância imagiológica.

c_lugar1

num2

num1

c_lugar1

num3

num4

c_lugar2

Figura 5.4: Identificação de entidades frase 4

Na frase da figura 5.3, a ocorrência relevante “6mm” é identificada como sendo pertencente ao complemento circunstancial de lugar “no lobo direito”, sendo este o único complemento circunstancial de lugar da frase e “6mm” não possui qualquer tipo de complemento circunstancial de lugar.

Na frase da figura 5.4, a ocorrência relevante “9mm” é identificada como sendo pertencente ao complemento circunstancial de lugar “no lobo esquerdo”, sendo que o complemento circunstancial de lugar pertencente à ocorrência relevante “9mm” é: “nesta região ístmica”.

5.3.1.3 Extração de descrições de condições clínicas

Para cada ocorrência relevante foi calculado o F1, sendo que o F1 da extração automática de descrições de condições clínicas representa uma média aritmética de todos os F1's calculados. Para a extração automática de descrições de condições clínicas, no primeiro conjunto, obtemos um F1 igual a 97.9%. O F1 está muito próximo de 1 o que significa que a extração automática de descrições de condições clínicas é eficaz. Em 259 das 266 ocorrências relevantes o F1, foi igual a 1, ou seja, o conjunto extraído automaticamente foi igual ao conjunto anotado.

5.3.1.4 Extração de termos relevantes

Para avaliar o sistema perante o primeiro conjunto de dados, identificamos as ocorrências relevantes, onde a extração de entidades falha, ou seja, onde foi extraído, erradamente, uma condição clínica, uma descrição de condição clínica ou um complemento circunstancial de lugar.

Para o primeiro conjunto de dados o sistema indica que 96.2% dos termos são corretamente extraídos, ou seja, 256 dos 266 termos foram corretamente extraídos.

5.3.2 Segundo conjunto de dados

O segundo conjunto de dados serve, exclusivamente, para validar o conjunto de regras geradas pelo primeiro conjunto. Este não foi usado para calibrar as regras geradas pelo primeiro conjunto.

5.3.2.1 Extração de condições clínicas

Resultados		Termo	
		Presente	Ausente
Extração Automática de Termo	Presente	235	1
	Ausente	0	21

Tabela 5.5: Resultados da extração automática condições clínicas, segundo conjunto

$$Accuracy = \frac{235+21}{235+1+0+21} = 0.996 \quad (5.7)$$

A extração automática de condições clínicas, para o segundo conjunto de dados, falhou apenas numa das 257 ocorrências relevantes. A *accuracy*, para o segundo conjunto de dados, é de 0.996, resultado idêntico ao do primeiro conjunto de dados. Fica reforçada a hipótese de que as regras produzidas pelo primeiro conjunto de dados para extração de condições clínicas são um bom conjunto de regras.

5.3.2.2 Extração de complementos circunstanciais de lugar

Resultados		Termo	
		Presente	Ausente
Extração Automática de Termo	Presente	202	4
	Ausente	2	49

Tabela 5.6: Resultados da extração automática de complementos circunstanciais de lugar, primeiro conjunto

$$Accuracy = \frac{202+49}{202+4+2+49} = 0.977 \quad (5.8)$$

Ao contrário do que acontece na extração de condições clínicas, onde os resultados do segundo conjunto de dados são melhores do que os do primeiro, na extração do complemento circunstancial de lugar os resultados do segundo conjunto de dados apresentam uma *accuracy* pior que a do primeiro conjunto. A *accuracy* da extração de complementos circunstanciais de lugar é de 0.977 e deve-se a uma única frase relevante, onde se encontram 4 das 6 ocorrências relevantes erradas. A frase é apresentada na Figura 5.5

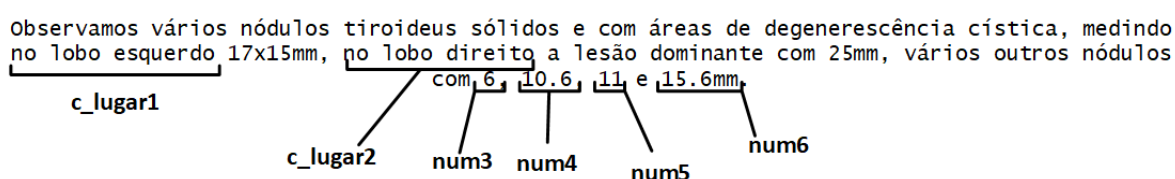


Figura 5.5: Identificação de entidades frase 5

Podemos perceber que para a extração do complemento circunstancial de lugar das ocorrências relevantes “6, 10.6, 11 e 15.6mm”, o sistema identifica as 4 ocorrências relevantes pertencentes ao “lobo direito”. No entanto, estas 4 ocorrências são implícitas e não permitem uma interpretação “humana” clara. Tanto podem referir-se a “no lobo esquerdo”, como “no lobo direito” ou a ambos. Sendo assim, ao assumir que se refere ao lobo direito, o sistema pode estar correto ou não.

Posto isto, se retirarmos as 4 ocorrências podemos perceber que os resultados da avaliação do segundo conjunto de dados são muito idênticos aos do primeiro conjunto de dados. Mais uma vez, fica reforçada a hipótese de que as regras utilizadas para a extração automática de complementos circunstanciais de lugar são um bom conjunto de regras.

5.3.2.3 Extração de descrições de condições clínicas

Utilizando o mesmo método de avaliação do primeiro conjunto de dados para o segundo conjunto de dados, obtemos um F1 igual a 98%, resultado em tudo idêntico ao primeiro conjunto. Sendo assim, fica reforçada a hipótese de que o conjunto de regras geradas pelo primeiro conjunto de dados para a extração de descrições de condições clínicas é um bom conjunto de regras.

5.3.2.4 Extração de termos relevantes

Da mesma forma que avaliamos para o primeiro conjunto de dados, também no segundo conjunto foi necessário saber em que ocorrências a extração de entidades falhou.

Para o segundo conjunto de dados isso acontece em 14 das 255 ocorrências relevantes, ou seja, o sistema extrai corretamente 94.5% dos termos relevantes.

5.3.3 Avaliação do sistema

Para avaliar o sistema tivemos em conta os dois conjuntos de dados.

Os resultados obtidos para a avaliação do sistema foram: para a extração de condições clínicas uma *accuracy* de 98.9%, para a extração de complementos circunstanciais de lugar uma *accuracy* de 98.5% e para as descrições de condições clínicas um F1 de 97.8%. O sistema indica que 95.4% dos termos relevantes são corretamente extraídos.

Capítulo 6

Conclusões

Esta dissertação apresenta um sistema de extração de termos relevantes, que visa, numa primeira fase, a identificação de ocorrências relevantes, numa segunda fase, a identificação das entidades e, por fim, a extração automática de termos relevantes com base nas entidades associadas às ocorrências relevantes. Sendo projetado em oito componentes diferentes: o pré-processamento de texto clínico, a identificação de possíveis ocorrências relevantes, a identificação de frases relevantes, o *POSTagging* e lematização de frases relevantes, a identificação de entidades, a construção de sequências das entidades identificadas, a associação das entidades às ocorrências relevantes e a extração de termos relevantes.

A avaliação do sistema foi feita recorrendo a dois conjuntos de dados previamente anotados, comparando os resultados da extração com as anotações. Para a extração de termos relevantes o sistema indica que 95.4% dos termos são corretamente extraídos, para a extração de condições clínicas obtivemos uma *accuracy* de 98.9%, para a extração de complementos circunstanciais de lugar uma *accuracy* de 98.5% e para as descrições de condições clínicas um F1 de 97.8%.

Perante estes resultados, podemos dizer que o presente sistema de extração automática de termos relevantes apresenta bons resultados para os relatórios da tiroide, concluindo que é possível usar esta metodologia para extrair automaticamente termos relevantes de relatórios da tiroide.

6.1 Discussão

Apesar dos bons resultados obtidos, a metodologia utilizada apresenta algumas limitações, principalmente, em três pontos: na identificação de entidades, na identificação de ocorrências relevantes e na associação das entidades às ocorrências relevantes.

A identificação de entidades necessita de um conhecimento prévio do léxico, o que por si, pode significar um problema, já que a identificação de entidades passa a ser sensível, por exemplo, a erros ortográficos, pois não identifica entidades que os possuam. Outro dos problemas na identificação de entidades é a identificação de complementos circunstanciais de lugar, esta está

extremamente dependente da gramática, pois existe a necessidade de identificar contrações da proposição “em”. Sendo que, identificar contrações da proposição “em” para identificar complementos circunstanciais de lugar não é um método infalível.

Para identificar uma possível ocorrência relevante, partimos do pressuposto que estas se encontram em milímetros (“mm”) e, mais uma vez, aqui, a identificação está dependente do léxico. Qualquer número que não apresente a unidade (tirando as enumerações, como referido em 4.2) está excluído à partida. Contudo, é preciso não esquecer que o sistema falha 7% das vezes quando uma possível ocorrência relevante é identificada erradamente, sendo que uma possível ocorrência relevante foi identificada erradamente pelo sistema 13% de todas as ocorrências relevantes analisadas. Refinar o processo de identificação de ocorrências relevantes poderá ser uma solução, não excluindo que a identificação de ocorrências relevantes é um processo crucial da presente metodologia para a extração dos termos relevantes.

Na associação das entidades às ocorrências relevantes como exemplificado na Figura 5.5, o sistema é pouco eficaz quando lida com texto implícito.

6.2 Trabalho futuro

A metodologia proposta nesta dissertação sugere um bom desempenho, mas poderá ser melhorada, começando por solucionar alguns dos problemas anteriormente descritos.

Além disso, é importante saber se é possível aplicar esta metodologia a relatórios de patologias diferentes com bons resultados, tendo também em consideração as limitações referidas acima. Para reconhecer entidades em relatórios de patologias diferentes será necessário recorrer a uma nova análise, nomeadamente no que diz respeito ao léxico da patologia abordada e analisar a possibilidade de se encontrar ocorrências nos termos relevantes de relatórios de patologias diferentes.

A metodologia apresentada poderá ser utilizada para treinar um modelo preditivo supervisionado, sendo o *input* a sequência e ocorrência relevante a ser extraída e o *output* a sequência relevante. Desta forma, será possível criar um modelo, baseado em métodos de *machine learning*, capaz de automatizar a obtenção de regras e consequentemente extrair automaticamente os termos relevantes. O processo de *machine learning* para obtenção automática de regras também poderia ser aplicado a relatórios de outras patologias.

Bibliografia

- [1] R. Braga, “Os Registos Clínicos e a Codificação,” *Revista Portuguesa de Medicina Geral e Familiar*, vol. 28, no. 3, pp. 155–156, 2012. [Online]. Available: http://www.scielo.gpeari.mctes.pt/scielo.php?script=sci_arttext&pid=S2182-51732012000300001&lng=pt&nrm=iso&tlng=pt
- [2] K. Dahlgren and E. Stabler, “Natural language understanding system,” 1998. [Online]. Available: <https://www.google.com/patents/US5794050>
- [3] R. Grishman and B. Sundheim, “Message Understanding Conference-6: A Brief History,” *Proceedings of the 16th conference on Computational linguistics*, vol. 1, pp. 466–471, 1996. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=992628.992709>
- [4] E. Marsh and C. Friedman, “Transporting the linguistic string project system from a medical to a Navy domain,” *ACM Transactions on Information Systems*, vol. 3, no. 2, pp. 121–140, 1985.
- [5] C. Friedman, S. B. Johnson, B. Forman, and J. Starren, “Architectural requirements for a multipurpose natural language processor in the clinical environment,” *Annual Symposium on Computer Applications in Medical Care*, vol. 19, pp. 347–351, 1995.
- [6] K. L. Moore, A. F. Dalley, and A. M. R. Agur, *Clinically oriented anatomy*. Lippincott Williams & Wilkins, 2013.
- [7] M. P. J. Vanderpump and W. M. G. Tunbridge, “The epidemiology of thyroid diseases,” *Werner and Ingbar’s the thyroid: a fundamental and clinical text*, pp. 398–406, 2005.
- [8] B. R. Haugen, E. K. Alexander, K. C. Bible, G. M. Doherty, S. J. Mandel, Y. E. Nikiforov, F. Pacini, G. W. Randolph, A. M. Sawka, M. Schlumberger, K. G. Schuff, S. I. Sherman, J. A. Sosa, D. L. Steward, R. M. Tuttle, and L. Wartofsky, “2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer,” *Thyroid*, vol. 26, no. 1, pp. 1–133, 2016. [Online]. Available: <http://online.liebertpub.com/doi/10.1089/thy.2015.0020>
- [9] B. Aschebrook-Kilfoy, R. B. Schechter, Y.-C. T. Shih, E. L. Kaplan, B. C.-H. Chiu, P. Angelos, and R. H. Grogan, “The Clinical and Economic Burden

- of a Sustained Increase in Thyroid Cancer Incidence,” *Cancer Epidemiology Biomarkers & Prevention*, vol. 22, no. 7, pp. 1252–1259, 2013. [Online]. Available: <http://cebp.aacrjournals.org/cgi/doi/10.1158/1055-9965.EPI-13-0242>
- [10] A. Hotho, A. Nürnberger, and G. Paaß, “A Brief Survey of Text Mining,” *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, vol. 20, pp. 19–62, 2005. [Online]. Available: <http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf>
- [11] R. Collobert and J. Weston, “A unified architecture for natural language processing,” *Proceedings of the 25th international conference on Machine learning - ICML '08*, vol. 20, no. 1, pp. 160–167, 2008. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1390177%5Cnhttp://portal.acm.org/citation.cfm?doid=1390156.1390177>
- [12] M. Rajman and R. Besançon, *Text Mining: Natural Language techniques and Text Mining applications*. Boston, MA: Springer US, 1998, pp. 50–64. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-35300-5{__}3
- [13] T. Güngör, “Part-of-Speech Tagging.” 2010. [Online]. Available: [ftp://nozdr.ru/biblio/kolxo3/Cs/CsNI/IndurkhyaN.,DameruF.J.\(eds.\)Handbookofnaturallanguageprocessing\(2ed.,CRC,2010\)\(ISBN9781420085921\)\(O\)\(692s\){__}CsNI{__}.pdf{#}page=231](ftp://nozdr.ru/biblio/kolxo3/Cs/CsNI/IndurkhyaN.,DameruF.J.(eds.)Handbookofnaturallanguageprocessing(2ed.,CRC,2010)(ISBN9781420085921)(O)(692s){__}CsNI{__}.pdf{#}page=231)
- [14] D. D. Palmer and M. A. Hearst, “Adaptive multilingual sentence boundary detection,” *Computational Linguistics*, vol. 23, no. 2, pp. 16–19, 1997.
- [15] R. Gaizauskas and Y. Wilks, *Information Extraction: Beyond Document Retrieval*, 1998, vol. 3, no. 2.
- [16] “Google.” [Online]. Available: <https://www.google.com>
- [17] D. Nadeau, “A survey of named entity recognition and classification,” *Linguisticae Investigationes*, no. 30, pp. 3–26., 2007. [Online]. Available: <http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>
- [18] C. Faucher and L. C. Jain, Eds., *Innovations in Intelligent Machines-4 - Recent Advances in Knowledge Engineering*, ser. Studies in Computational Intelligence. Springer, 2014, vol. 514. [Online]. Available: <http://dx.doi.org/10.1007/978-3-319-01866-9>
- [19] H. Schmid, “Probabilistic Part-of-Speech Tagging Using Decision Trees,” *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49, 1994.
- [20] “portuguese-finegrained-par-linux-3.2-utf8.bin.gzr.” [Online]. Available: www.cis.uni-muenchen.de/{~}schmid/tools/TreeTagger/data/portuguese-finegrained-par-linux-3.2-utf8.bin.gz
- [21] K. Hornik, “Package ‘openNLP’,” 2016. [Online]. Available: <https://cran.r-project.org/web/packages/openNLP/openNLP.pdf>

- [22] “openNLPmodels.pt_1.5-2.” [Online]. Available: https://www.datacube.wu.ac.at/src/contrib/openNLPmodels.pt{__}1.5-2.tar.gz
- [23] A. D. Shah, C. Martinez, and H. Hemingway, “The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records,” *BMC Medical Informatics and Decision Making*, vol. 12, p. 88, 2012.
- [24] G. Karystianis, T. Sheppard, W. G. Dixon, and G. Nenadic, “Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database,” *BMC Med Inform Decis Mak*, vol. 16, no. 1, p. 18, 2016. [Online]. Available: [{%}5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC4748480/pdf/12911{__}2016{__}Article{__}255.pdf">http://www.ncbi.nlm.nih.gov/pubmed/26860263{%}5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC4748480/pdf/12911{__}2016{__}Article{__}255.pdf](http://www.ncbi.nlm.nih.gov/pubmed/26860263)
- [25] S. Hassanpour and C. P. Langlotz, “Information extraction from multi-institutional radiology reports,” *Artificial Intelligence in Medicine*, vol. 66, pp. 29–39, 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.artmed.2015.09.007>
- [26] H. D. Tolentino, M. D. Matters, W. Walop, B. Law, W. Tong, F. Liu, P. Fontelo, K. Kohl, and D. C. Payne, “A UMLS-based spell checker for natural language processing in vaccine safety,” *BMC medical informatics and decision making*, vol. 7, p. 3, 2007. [Online]. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1805499{&}tool=pmcentrez{&}rendertype=abstract>
- [27] J. Patrick and C. Street, “Text Mining in Clinical Domain : Dealing with Noise,” pp. 549–558, 2015. [Online]. Available: <http://www.kdd.org/kdd2016/papers/files/adp0478-nguyenA.pdf>
- [28] X. Li, S. Qing, H. Zhang, T. Wang, and H. Yang, “Kernel methods for word sense disambiguation,” *Artif. Intell. Rev.*, vol. 46, no. 1, pp. 41–58, 2016. [Online]. Available: <http://dx.doi.org/10.1007/s10462-015-9455-5>
- [29] J. P. Ferraro, H. Daumé, S. L. Duvall, W. W. Chapman, H. Harkema, and P. J. Haug, “Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation.” *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, no. 5, pp. 931–9, 2013. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/23486109>
- [30] A. B. Clegg and A. J. Shepherd, “Benchmarking natural-language parsers for biological applications using dependency graphs.” *BMC bioinformatics*, vol. 8, p. 24, 2007.
- [31] Y. Kodratoff, “Knowledge discovery in texts: A definition, and applications,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1609, pp. 16–29, 1999.
- [32] A. Baneyx, J. Charlet, and M.-C. Jaulent, “Building an ontology of pulmonary diseases with natural language processing tools using textual corpora,” *I. J. Medical Informatics*, vol. 76, no. 2-3, pp. 208–215, 2007. [Online]. Available: <http://dx.doi.org/10.1016/j.ijmedinf.2006.05.031>

- [33] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries." *Journal of biomedical informatics*, vol. 34, no. 5, pp. 301–10, 2001. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1532046401910299>
- [34] W. Sun, A. Rumshisky, and O. Uzuner, "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge." *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, no. 5, pp. 806–13, 2013. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/23564629>
- [35] P. Chen, A. Barrera, and C. Rhodes, "Semantic analysis of free text and its application on automatically assigning ICD-9-CM codes to patient records," *Proceedings of the 9th IEEE International Conference on Cognitive Informatics, ICCI 2010*, pp. 68–74, 2010.
- [36] "Ministério da Saúde." [Online]. Available: http://www.portalcodgdh.0min-saude.pt/index.php/Codifica{ç}{~{a}}o{_{}}cl{i}nica
- [37] N. Sager, C. Friedman, E. Chi, C. Macleod, S. Chen, and S. Johnson, "The analysis and processing of clinical narrative," *Medinfo*, vol. 86, pp. 1101–1105, 1986.
- [38] P. J. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha, and S. M. Huff, "Experience with a mixed semantic/syntactic parser." *Proceedings of the Annual Symposium on Computer Application in Medical Care*, p. 284, 1995.
- [39] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association : JAMIA*, vol. 17, no. 5, pp. 507–513, 2010. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995668/{%}5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995668/pdf/amiajnl1560.pdf>
- [40] J. Friedlin and C. J. McDonald, "A Natural Language Processing System to Extract and Code Concepts Relating to Congestive Heart Failure from Chest Radiology Reports," in *{AMIA} 2006, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 11-15, 2006*. AMIA, 2006. [Online]. Available: <http://knowledge.amia.org/amia-55142-a2006a-1.620145/t-001-1.623243/f-001-1.623244/a-054-1.623586/a-055-1.623583>
- [41] J. Baldridge, "The opennlp project," 2005. [Online]. Available: <http://opennlp.apache.org/index>
- [42] C. Lindberg, "The Unified Medical Language System (UMLS) of the National Library of Medicine," *Journal (American Medical Record Association)*, vol. 61, no. 5, pp. 40–42, 1990. [Online]. Available: <http://europepmc.org/abstract/MED/10104531>
- [43] National Library of Medicine, "UMLS® Reference Manual - NCBI Bookshelf," no. Md, 2009. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK9676/>

-
- [44] L. Ferreira and S. Cunha, “Medical Information Extraction in European Portuguese,” 2006.
 - [45] M. R. Rocha, *Gramática de Português - Ensino Secundário*. Porto Editora, 2016.
 - [46] C. Amorim and C. Sousa, *Gramática do Português*. Areal Editores, 2016.
 - [47] R. Srikant and R. Agrawal, *Mining sequential patterns: Generalizations and performance improvements*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 1–17. [Online]. Available: <https://doi.org/10.1007/BFb0014140>

Anexo

Neste anexo são enumeradas as listas de palavras, usadas no reconhecimento de entidades, das condições clínicas, das descrições das condições clínicas e das exceções. Estas listas são usadas nos algoritmos 4.7, 4.8 e 4.9.

1. nódulos	10. pseudonodular	19. perinodular
2. nodularidades	11. nodularidade	20. macronódulos
3. nódulo	12. nodulariformes	21. micronodularidades
4. nodulares	13. pseudonodulares	22. nodulariforme
5. micronódulos	14. pseudomicronodulares	23. noduloso
6. multinodular	15. nodulações	24. micronodulares
7. micronódulo	16. pseudonódulo	25. pseumicronodulares
8. nodular	17. intranodular	26. nodulos
9. micronodular	18. pseudonódulos	

Figura A.1: Palavras da família “nódulo”

1. quisto	3. microquistos	5. cisto	7. microcistos
2. quistos	4. microquisto	6. cistos	

Figura A.2: Palavras da família “quisto” e “cisto”

1. colóide	2. coloide	3. colóides	4. coloides
------------	------------	-------------	-------------

Figura A.3: Palavras da família “cóloide”

1. sólido	2. sólidas	3. sólidos	4. sólida
-----------	------------	------------	-----------

Figura A.4: Palavras da família “sólido”

1. mista	2. misto	3. mistas	4. mistos
----------	----------	-----------	-----------

Figura A.5: Palavras da família “misto”

1. hipervascular	4. vascularizado	7. vasculares
2. vascularização	5. hipervascularização	8. vascular
3. vascularizada	6. hipervascularizado	9. vascularizados

Figura A.6: Palavras da família “vascularização”

1. calcificações	3. microcalcificações	5. calcificados	7. calcificada
2. calcificado	4. calcificação	6. calcificadas	8. microcalcificação

Figura A.7: Palavras da família “calcificação”

1. hipoecóicos	7. hipoecoica	13. isoecóica	19. hiperecoicas
2. hipoecóico	8. hipoecoicas	14. isoecoica	20. hiperecóicos
3. hipoecóicas	9. hipoecóicae	15. isoecoicas	21. anecoicas
4. hipoecoicos	10. isoecoico	16. isoecóicos	22. hiperecóica
5. hipoecóica	11. isoecóicas	17. isoecocio	23. hiperecoicos
6. hipoecoico	12. isoecóico	18. isoecoicos	24. hiperecóico

Figura A.8: Palavras da família “ecóicos”

1. hipoecogénico	14. hipoecogenicos	27. isoecogéneo	40. isoecogenea
2. hipoecogénica	15. hipoecógena	28. ecogenicidade	41. ecogeno
3. hipoecogénicos	16. hipoecogéneas	29. hiperecogénico	42. isoecogéneos
4. hipoecogénicas	17. hipoecogéneos	30. hiperecogénicos	43. isoecogéneas
5. hipoecogénea	18. hipoecogéneo	31. hiperecogénicas	44. ecogénica
6. hipoecogéneo	19. hipoecogéneos	32. isoecogénea	45. issoecogénica
7. hipoecogeno	20. hipoecocoica	33. ecogéneos	46. isoecogenico
8. hipoecogenos	21. hipoecógenica	34. ecogénico	47. anecogénica
9. hipoecógeno	22. hipoecogenica	35. isoecogénica	48. anecogénicas
10. hipoecogenicidade	23. ecogénicos	36. isoecogeno	49. ecogéneo
11. hipoecógenos	24. isoecogénicas	37. ecogenecidade	50. isoecogenas
12. hipoecogéno	25. hiperecogénica	38. isoecogénicos	51. ecogenos
13. hipoecogenicidades	26. isoecogénico	39. ecogénicas	52. isoecogena

Figura A.9: Palavras da família “écogenico”

1. cística	5. císticos	9. quísticos	13. quistico
2. císticas	6. cisticas	10. quística	14. pseudoquístico
3. cístico	7. quístico	11. quisticos	15. intraquística
4. cisticos	8. quísticas	12. pseudoquistico	

Figura A.10: Palavras da família “quístico”/“cístico”