TÉCNICO LISBOA

Identificação e Classificação de Entidades Mencionadas em Documentos da Marinha

Gonçalo Azevedo Rodrigo

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Prof. Nuno João Neves Mamede

Coorientadores: Prof. Jorge Manuel Evangelista Baptista

Júri

Presidente: Prof. Luís Manuel Antunes Veiga

Orientador: Prof. Nuno João Neves Mamede

Vogal: Prof. Ricardo Daniel Santos Faro Marques Ribeiro

Junho 2020

Agradecimentos

Gostaria de agradecer a dedicação, incentivo e colaboração dos meus orientadores - Professor Nuno Mamede e co-orientador - Professor Jorge Baptista.

Quero agradecer ao Filipe Reis pela disponibilidade e pelo apoio dado. Foi importante no esclarecimento de dúvidas acerca da Marinha Portuguesa.

Um obrigado especial à Melanie Bernardo pela motivação e confiança. Foi essencial no incentivo para a conclusão desta etapa. Também gostaria de agradecer aos meus pais, avós e irmãos, sem eles isto não seria possível.

Aos meus colegas Alexandre Machado, André Xiang, Bruno Lopes, Filipe Azevedo, Inês Leite, Martim Zanatti e Pedro Santos pelos momentos de entreajuda, confraternidade e animação. Todos os momentos contribuíram para uma amizade que permanecerá.

Por fim, gostaria de agradecer a todos os meus professores pela contribuição no meu desenvolvimento cognitivo. Também gostaria de agradecer a todos os membros do (L^2F) que de uma forma ou de outra contribuíram para a realização deste trabalho.

Resumo

A informação de uma organização consiste maioritariamente em informação não estruturada. De forma a transformá-la em informação útil, foram desenvolvidas técnicas e ferramentas para a Extração de Informação (EI). Uma das tarefas da EI é o Reconhecimento e Classificação de Entidades Mencionadas (RCEM). O conceito de entidade mencionada foi inicialmente proposto pela Conferência MUC-6, em 1996. Desde então, múltiplas técnicas foram desenvolvidas para extrair entidades de diversos tipos de textos e para várias línguas. Mesmo assim, na comunidade de investigadores, o interesse para desenvolver novas abordagens para identificar e classificar entidades mencionadas mantêm-se, visto que esta operação permite extrair conhecimento do texto. Neste projeto, realizamos o tratamento dos documentos da Marinha Portuguesa, de forma a produzir um *Corpus*. Usando o *Corpus*, também, testamos a tarefa de RCEM da nossa cadeia de processamento de Língua Natural.

Abstract

An organization's information consists mostly of unstructured information. To transform it into useful information, techniques and tools for Information Extraction (IE) were developed. One of IE's tasks is Named-Entity Recognition and Classification (NERC). The named-entity concept was initially proposed by the MUC-6 Conference in 1996. Since then, multiple techniques have been developed to extract entities from different types of texts and for several languages. Even so, in the community of researchers, the interest to develop new approaches to identify and classify mentioned entities remains, since this operation allows to extract knowledge from the text. In this project, we carry out the treatment of Portuguese Navy documents, to produce a *Corpus*. Using *Corpus*, we also tested the NERC task of our Natural Language Processing chain.

Palavras-Chave

Palavras-Chave

Processamento de Língua Natural (PLN)

Reconhecimento e Classificação de Entidades Mencionadas (REM)

Keywords

Natural Language Processing (NLP)

Named Entity Recognition and Classification (NERC)

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

IDC	**I**nternational **D**ata **C**orporation
INESC-ID	**In**stituto **E**ngenharia e de **S**istemas e **C**omputadores - **I**nvestigação e **D**esenvolvimento
LexMan	**Lex**ical **M**orfological **An**alizer
L^{2}F	**L**aboratório de Sistemas de **L**íngua **F**alada
MARv	**M**orphosyntactic **A**mbiguity **R**esol**v**er
PLN	**P**rocessamento de **L**íngua **N**atural
POS	**P**art **o**f **S**peech - Classe de uma palavra
REM	**R**econhecimento de **E**ntidades **M**encionadas
RuDriCo	**Ru**le-**Dri**ven **Co**nverter
STRING	**St**atistical and **R**ule-based natural language process**ing**
XIP	**X**erox **I**ncremental **P**arser

Capítulo 1

Introdução

Os dados de uma organização podem dividir-se em duas categorias: estruturados e não estruturados. Os dados não estruturados incluem: ficheiros, documentação, emails, planos de projectos, manuais de produtos, páginas da *WEB*, etc., e são criados em diferentes suportes e formatos. Em 1998, Merrill Lynch apresentou uma regra que dizia que 80%-90% de todos os dados de uma organização não são estruturados [32]. A mesma percentagem prevalece até à atualidade [22]. Além disso, em 2017, a IDC previu que haverá 10 vezes mais dados em 2025 [27].

A *explosão* de informação exigiu a procura de métodos mais eficientes para o processamento de documentos não estruturados. O ramo da ciência da computação que se preocupa com a resolução deste problema, assim como a interpretação e geração automática de linguagem humana é o Processamento de Língua Natural (PLN)[1].

A extração de informação é uma tarefa do PLN, que, por sua vez, tem uma sub-tarefa designada Reconhecimento e Classificação de Entidades Mencionadas (RCEM)[2]. Em 1996, na conferência MUC-6 [14] foi definido o conceito de *Entidade Mencionada (EM)*. Este conceito surgiu após o reconhecimento da *unidade de informação* como elemento fundamental para a tarefa de extração de informação. A EM é uma expressão linguística usada para designar objetos do mundo real (pessoas, locais, organizações, etc), geralmente correspondentes a nomes próprios. Para além dos nomes próprios, outro tipo de expressões são usados no RCEM, nomeadamente as expressões temporais (datas, períodos, efemérides, etc) e as expressões numéricas (número de contribuinte, matrículas de carros, etc). Além disso, o tipo de Entidade Mencionada depende do domínio de interesse. Por exemplo, no domínio Militar, algumas entidades relevantes são: a patente, o órgão, a missão, entre outros; por outro lado, no domínio geral, as entidades relevantes são as seguintes: a pessoa, a localização, a organização, os valores numéricos, as expressões temporais, entre outros.

A tarefa de identificar e classificar corretamente este tipo de expressões é essencial para uma análise semântica do texto e facilitar o processamento sintático subsequente. O RCEM também auxilia em algumas tarefas do PLN, tais como Sumarização Automática de Texto [23], Tradução Automática [3], Recuperação de Informação [16], Sistema de Pergunta-Resposta [26] e Reconhecimento de Fala.

As principais contribuições para o RCEM provêm das técnicas e ferramentas desenvolvidas para vários eventos

[1] Em inglês *Natural Language Processing (NLP). Existem certos termos específicos que não têm tradução direta para o português, de forma a melhorar a compreensão do texto foi a adicionado no rodapé os termos em inglês.*

[2] Do inglês *Named Entity Recognition and Classification (NERC).*

científicos, tais como: o *Information Retrieval and Extraction (IREX)* [31], a *Conference on Natural Language Learning 2002 (CONLL 2002)* [33] e *2003 (CONLL 2003)* [34], o *Automatic Content Extraction (ACE) Program* [11], e, para o Português, a *Avaliação de Reconhecimento de Entidades Mencionadas* (*HAREM*) [30].

Neste projeto, a cadeia de PLN usada é a STRING[3] [15]. Esta cadeia é analisada detalhadamente no Capítulo 2. A STRING realiza tarefas básicas de processamento de texto em português, tais como a segmentação de texto e sua atomização[4], anotação morfossintática [5], desambiguação morfossintática, análise sintática do texto em constituintes nucleares[6] e extração de dependências.

A STRING está organizada em quatro módulos. O *LexMan* é o primeiro módulo, recebe o texto a processar e realiza a sua segmentação, definindo os segmentos que compõem o texto. O *LexMan* é um anotador morfossintático que recebe o resultado da segmentação como *input*, associa todas as categorias gramaticais possíveis a cada segmento e agrupa os segmentos em frases. O próximo módulo, designado *RubriCo2* é um desambiguador morfossintático baseado em regras e pode alterar a segmentação do módulo anterior, por exemplo, agrupar segmentos que formam palavras compostas. O terceiro módulo é o *MARv4*, um desambiguador morfossintático estatístico, que recebe o resultado do *RuDriCo2* e selecciona para cada segmento a categoria gramatical mais provável. Por fim, o último módulo a aplicar é o *XIP*, responsável pela análise sintática. Além destas tarefas básicas, esta cadeia de PLN é ainda capaz de realizar Reconhecimento de Entidades Mencionadas, Recuperação de Informação, Resolução de Anáforas e outras tarefas de PLN.

1.1 Problema

A STRING foi inicialmente desenvolvida apenas para processar texto de natureza geral (*e.g.* texto jornalístico). O *Corpus* utilizado neste trabalho consiste em correspondência da Marinha, um domínio textual particular, logo, existem termos compostos, entidades mencionadas e eventos que a STRING não identificou e classificou incorretamente.

Apesar de o RCEM ser considerado, geralmente, uma tarefa com o objetivo atingido devido às suas elevadas taxas de desempenho nas conferências científicas. Na verdade, estas avaliações usam um conjunto limitado de tipos de EM, que raramente se alteram ao longo dos anos. Além disso, usam *Corpus* de dimensões reduzidas, essencialmente quando comparados com outras áreas de Extração de Informação. Estes fatores conduzem a um sobreajustamento[7] das ferramentas e, consequentemente, a uma limitação da evolução na área [18]. Ou seja, o RCEM talvez seja um desafio resolvido para o domínio textual geral, contudo existe um défice de RCEM para os domínios de interesse específico.

1.2 Objectivos

[3] http://string.l2f.inesc-id.pt/ (última visita a 29/05/2020)
[4] Em inglês *tokenization*
[5] Em inglês *POS(Part-of-Speech) tagging*
[6] Em inglês *chunking*
[7] Em inglês *overfitting*

O objetivo principal deste projeto é adaptar um sistema de PLN, em particular o seu módulo de RCEM, inicialmente desenvolvido para processar textos de natureza geral (e.g. texto jornalístico), para um domínio textual particular, a correspondência oficial da Marinha Portuguesa.

Com o processamento do *Corpus* na cadeia STRING, pretende-se aumentar o número de entidades mencionadas e eventos identificados e classificados, para tal iremos realizar as seguintes tarefas:

- Constituição de um *Corpus* anotado de um domíno específico;
- Identificação de novos termos compostos;
- Identificação e Classificação de novas entidades mencionadas;
- Identificação de novos eventos relativos às Forças Armadas Portuguesas.

1.3 Contributos esperados

Tendo como ponto de partida a informação não estruturada, isto é, a documentação oficial da Marinha Portuguesa, iremos tratá-la, de forma a que a cadeia de PLN consiga processar estes documentos. De seguida, serão adicionadas regras à STRING para reconhecer as novas entidades mencionadas e classificá-las de acordo com o seu tipo no documento. Este projeto também pretende contribuir para o progresso na tarefa de RCEM. Para tal, iremos utilizar um domínio textual particular, contrariamente ao domínio textual geral utilizado nas conferências e fóruns de RCEM.

Este projeto é o primeiro passo de dois para realizar a distribuição automática de documentos da Marinha. O segundo passo consiste na classificação dos documentos utilizando as técnicas de aprendizagem automática.

1.4 Estrutura do Documento

No Capítulo 2 é descrita a cadeia de processamento de língua natural, a STRING. Esta cadeia é dividida em três etapas: o pré-processamento, a desambiguação morfossintática e a análise sintática. O Reconhecimento e Classificação das Entidades Mencionadas é aprofundado neste Capítulo.

O Capítulo 3 descreve o *Corpus* utilizado: primeiro, descrevemos a instituição na qual os documentos tiveram origem; de seguida, é realizada uma análise estrutural dos documentos; por fim, é apresentado o processo utilizado para transformar os documentos com informação não estruturada em textos processáveis pela STRING, bem como, as dificuldades encontradas.

A implementação e os procedimentos são descritos no Capítulo 4. Este capítulo apresenta o procedimento utilizado para a anotação das Entidade Mencionadas na Coleção Dourada, assim como as directivas da STRING utilizadas.

No Capítulo 5, é descrita a metodologia usada na avaliação do projeto e os resultados obtidos.

Por fim, no Capitulo 6 são apresentados algumas ideias para trabalhos futuros e as conclusões a retirar deste projeto.

Capítulo 2

Arquitectura

Neste capítulo é apresentada a arquitectura geral do sistema PLN, na qual a tarefa de reconhecimento de entidades mencionadas está incluída. Na secção 2.1, a STRING é descrita de uma forma geral. De seguida, na secção 2.2 é apresentada uma descrição detalhada do sistema XIP, cobrindo as regras de depêndencia, assim como o léxico personalizado e as gramáticas locais.

2.1 Cadeia de Processamento

A STRING é uma cadeia de PLN para o português baseada em regras e métodos estatísticos. Esta ferramenta foi desenvolvida pelo Laboratório de Língua Falada (L^2F) do INESC-ID em Lisboa [15]. Como mostra a Figura 2.1, a cadeia é constituída por quatro módulos e o processamento divide-se em três etapas:

- Pré-processamento;
- Desambiguação (guiado por regras e métodos estatísticos);
- Análise sintáctica.

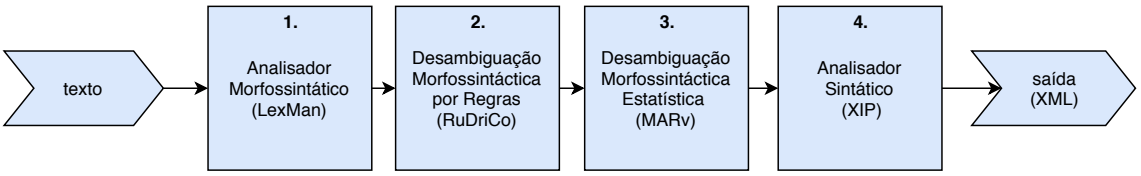

Figura 2.1: Cadeia de Processamento STRING.

2.1.1 Pré-processamento

O pré-processamento corresponde ao Analisador Morfossintático da Figura 2.1. Este módulo é dividido em três etapas. A primeira etapa é a segmentação, responsável principalmente por dividir a entrada em segmentos, também

conhecidos como *tokens*. Devido aos segmentos individuais serem chamados de *tokens*, esta etapa também pode ser chamada de *tokenizer*. Como exemplo, considere-se a frase *O Sr. João foi à Índia.* como entrada para a etapa de segmentação. Neste caso, a saída seria a representação da Figura 2.2.

```
word[0]: |O|
word[1]: |Sr.|
word[2]: |João|
word[3]: |foi|
word[4]: |à|
word[5]: |Índia|
word[6]: |.|
```

Figura 2.2: O *output* após a etapa de segmentação da frase.

Para além disso, esta etapa também é responsável pela identificação de sinais de pontuação e símbolos, assim como expressões alfanuméricas algumas das quais correspondem a determinados tipos de Entidades Mencionadas, tais como:

- Números ordinais (*e.g.* 2º,10º);
- Números com "."e ","(*e.g.* 13.570,95);
- Números inteiros (*e.g.* 1234);
- Números romanos (*e.g. LI*, *MMM*, *XI*);
- Adereços *IP* e *HTTP*;
- Adereços de *mail*;
- Abreviações com "."(*e.g.* "a.c", "V.Exa.","Sr.");
- Palavras (*e.g.* África do Sul).

Ainda nesta etapa, os números por extenso são identificados, por exemplo *oito mil e quinhentos e quarenta e três* é etiquetado como número [25]. De seguida, é realizada a anotação morfossintática[1], que resulta a associação de uma ou mais categorias gramaticais a cada segmento. O sistema de anotação morfossintático utiliza um conjunto de doze categorias gramaticais: nome, verbo, adjetivo, pronome, artigo, advérbio, preposição, conjunção, numeral, interjeição, símbolo e pontuação; e cada etiqueta é composta por dez campos: Categoria (CAT), Subcategoria (SCT), modo (MOD), tempo (TEN), pessoa (PER), número (NUM), género (GEN), grau (DEG), caso (CAS) e formação (FOR).

Considere-se de novo a frase de exemplo apresentada acima. Nesta etapa, o *output* após a anotação morfossintática seria a representação da Figura 2.3.

Na Figura 2.3, podemos constatar que é associada, pelo menos, uma anotação morfossintática para cada segmento. Por exemplo, *João* e *Índia* são associados, respectivamente, *NP...smn.==* e *NP...sfn.==*, o que significa que ambos são nomes próprios com número singular e, respectivamente, género masculino e feminino. No entanto, existem segmentos ambíguos, ou seja, que pertencem a mais do que uma categoria gramatical ou correspondem

[1] *Part-of-Speech (POS) tagging*

```
word[0]: |O|        POS-> [eu] Pp..3sm.a==
                          [o] Td...sm...=
word[1]: |Sr.|      POS-> [senhor] Nc...smn.=a
word[2]: |João|     POS-> [João] Np...smn.==
word[3]: |foi|      POS-> [ir] V.is3s=...=
                          [ser] V.is3s=...=
word[4]: |à|        POS-> [ao] S....sf...=
word[5]: |Índia|    POS-> [Índia] Np...sfn.==
word[6]: |.|        POS-> [.] O..........
```

Figura 2.3: O processo de anotação morfossintático da frase.

a flexões homógradas de lemas diferentes. Por exemplo, o segmento *foi*, que é um verbo, pode ter como lema o verbo ir ou ser. Esta anotação é realizada pelo módulo *LexMan* [20].

A última etapa do pré-processamento é a divisão do texto em frases. De forma a segmentar as frases, o sistema baseia-se, sobretudo nas frases que acabam com ".", "!"ou "?". Não obstante, existem duas exceções a esta regra:

- Todas as abreviaturas registadas (e.g. N.A.S.A.);
- Se houver um símbolo ou letra minúscula depois dos seguintes sinais de pontuação: "»", ")", "]", "}".

Finalmente, a saída é convertida em XML por um conversor que faz a ligação entre o primeiro e o segundo módulo. A fase seguinte consiste na Desambiguação Morfossintática baseada em regras.

2.1.2 Desambiguação

A próxima etapa na cadeia de processamento é o processo de desambiguação, no qual compreende dois passos:

- Desambiguação morfossintática baseada em regras, executada pelo RuDriCo [9, 10];
- Desambiguação estatística, executada pelo MARv [29].

Desambiguação baseada em regras

O principal objetivo do RuDriCo, de acordo com Diniz [9], é fornecer um ajuste nos resultados produzidos pelo analisador morfossintático para as necessidades específicas de cada analisador sintático [2]. De forma a atingir este objetivo, o RuDriCo modifica a segmentação anteriormente realizada pelo *LexMan*. Por exemplo, pode unir dois *tokens* num único, expressões como *ex-* e *namorada* em *ex-namorada*; ou pelo contrário, expandir uma expressão, uma contração *à* em dois segmentos, *a*. Na Figura 2.4, pode-se ver um exemplo de expansão da palavra *à*. Esta modificação depende do que o *parser* poderá necessitar como *input*.

[2] *parser*

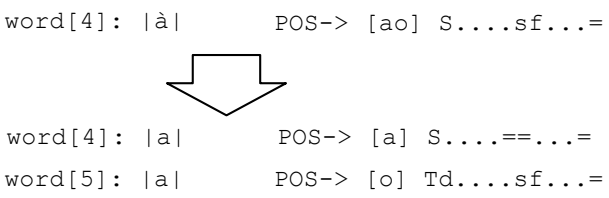

Figura 2.4: Expansão de um segmento em dois na frase da Figura 2.2.

Alterar a segmentação é também útil para as tarefas de reconhecimento de números e datas. Esta alteração é realizada através de regras declarativas, que se baseiam no conceito de emparelhamento de padrões. Esta ferramenta também pode ser usada para resolver ambiguidades morfossintáticas. Para além das tarefas mencionadas, o RuDriCo também corrige alguns *outputs* do LexMan e modifica o lema dos pronomes, advérbios, artigos,etc.

Atualmente é usado o RuDriCo 2.0, que resulta do melhoramentos de uma versão anterior do RuDriCo[9].

Por fim, o output é convertido para o terceiro módulo da Figura 2.1, a Desambiguação Morfossintática Estatística.

Desambiguação Estatística

O principal objectivo do MARv [28] é analisar as anotações morfossintáticas atribuídas a cada *token* no passo anterior da cadeia de processamento, e de seguida escolher a anotação mais provável para cada um. De forma a atingir este objectivo é usado um modelo estatístico conhecido como Modelo Escondido de Markov [3](MEM). Este modelo permite calcular a probabilidade do conjunto de estados ocultos dado um conjunto de estados observados. Nesta etapa, o MEM é utilizado para determinar as anotações morfossintáticas (estado oculto) das palavras (valores observados) na frase. Primeiro, é necessário atribuir uma probabilidade a cada sequência possível de estados ocultos, e, de seguida, escolher a sequência mais provável de anotações morfossintáticas na frase, usando o algoritmo de Viterbi [35]. Na Figura 2.5 é possível ver a desambiguação do lema verbal usando este método estatístico numa palavra pertencente à frase exemplo.

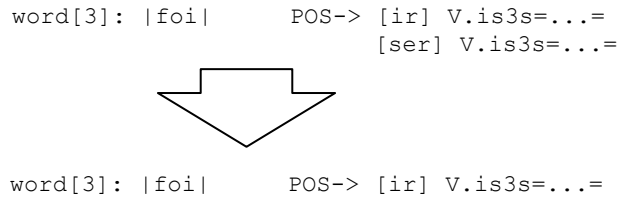

Figura 2.5: Desambiguação entre as duas anotações através de métodos estatísticos.

Atualmente, a cadeia de processamento utiliza o MARv 3.0, mais rápido que a versão anterior e guarda as anotações antigas [28].

Finalmente, o *output* é convertido entre o terceiro módulo e o último módulo, a Análise Sintática, pelo último conversor, como está representado na Figura 2.1.

[3] *Hidden Markov Model (HMM)*

2.1.3 Análise Sintática

A terceira e última etapa da cadeia de processamento é a análise sintáctica realizada pelo XIP da Xerox. Nesta etapa acontece a identificação e classificação de EMs, logo o principal trabalho deste projecto ocorre neste módulo. O XIP é um compilador de regras que adiciona informação linguística (sintáctica e semântica) ao retorno da anotação morfossintática. Esta ferramenta acede ao contexto circulante, assim como permite representar e manipular várias características linguísticas. O sistema é independente da língua, sendo que novas regras podem ser criadas incrementalmente sobre as existentes. O XIP também usa funcionalidades de *parsing* para dividir o texto em sintagmas nucleares [4] como é o caso do grupo nominal (GN) e grupo verbal (GV). De seguida, são extraída as relações sintácticas entre as cabeças dos *chunks*. Estas relações representam a principal funcionalidade de *parsing* entre dependências sintáticas (E.g. Sujeito-Complemento Directo, etc.), mas também inclui dependências auxiliares entre diferentes *chunks* e palavras, por exemplo a ligação entre os segmentos verbais e as palavras auxiliares [4].

Para além destes módulos, o texto pode ser sujeito a processamento por outros módulos, como é o caso do módulo que resolve as anáforas [24] e a normalização de expressões temporais [19]. Neste processamento da língua natural resulta um ficheiro XML que contém o resultado do XIP.

Para concluir esta secção, é importante mencionar que os vários módulos da cadeia de processamento podem ser parametrizados através de:

1) **LexMan:** Listas de abreviaturas;

2) **RuDriCo:** Dicionários de palavras;

3) **MARv:** Listas de regras de desambiguação e de relaxamento;

4) **XIP:** Gramáticas locais e léxicos.

2.2 Estrutura das Regras e Léxicos

No XIP, as categorias e subcategorias de cada entidade mencionada são representadas através de traços. Com os traços, pode-se introduzir informação lexical através de ficheiros de léxico. Além disso, as regras de identificação e classificação de EM são definidas nas gramáticas locais. Nesta fase do PLN, também é possível aplicar regras de desambiguação morfossintáticas.

A unidade de representação dos dados no XIP é o nó. O nó tem uma categoria, um conjunto de traços (pares de atributo-valor) e nós *irmãos*. Todos os traços e valores possíveis têm de ser declarados explicitamente, com excepção dos seguintes traços geridos pelo sistema: `lemma`, `surface`, `maj` e `toutmaj`.

Os traços `lemma` e `surface`, cujo valor são cadeia de caracteres, correspondem, respectivamente, ao lema da unidade linguística e à forma de superfície da unidade linguística; `maj` e `toutmaj` são traços booleanos que indicam, respectivamente, se a forma de superfície começa por maiúscula, ou se a forma de superfície é em maiúsculas. Todos os outros traços associados ao léxico são traços sintácticos ou semânticos e são definidos na gramática.

[4] *Traduzido da palavra inglesa chunk. Mais precisamente, por sintagma nuclear consideramos um grupo sintáctico, não recursivo, cujo limite direito corresponde à cabeça sintáctica dum sintagma tradicional.*

Por exemplo, o nó abaixo da Figura 2.6 representa o nome *João* e tem alguns traços que são usados para expressar as suas propriedades. Neste caso, os traços têm os seguintes significados: *João* é um nome que representa um humano (traço *human*) do género masculino (traço *masc*); o nó também tem atributos que descrevem o seu número (singular, atributo *sg*) e a primeira letra da palavra, neste caso está em letra maiúscula (atributo *maj*).

```
João: noun[human, individual, proper, firstname, people, sg, masc, maj]
```

Figura 2.6: Traços do nome próprio *João*

Todas as categorias e atributos dos nós têm de ser declaradas nos ficheiros de declaração, assim como cada atributo têm de ser declarado no seu domínio de valores possíveis. Estes ficheiros são uma parte importante do XIP, visto que descrevem as propriedades das unidades de representação dos dados, os nós. Os atributos não existem isolados, têm de estar sempre associados a um valor, daí serem chamados de pares atributo-valor.

Além disso, os atributos podem ser instanciados (operador =), testados (operador :), ou eliminados (operador =~) dentro de todos os tipos de regras. Enquanto a instanciação e a eliminação alteram ou removem valores dos atributos, o teste consiste em verificar se um valor específico está associado a um atributo. A Tabela 2.1 exemplifica as operações que são possiveis realizar sobre os pares atributo-valor.

Tipo	Exemplo	Explicação
Instanciado	[gender = fem]	O valor *fem* é associado ao atributo *gender*.
Testado	[gender : fem]	O atributo *gender* têm o valor *fem*?
	[gender :~]	O atributo *gender* não deve ser instanciado no nó.
	[gender :~fem]	O atributo *gender* não deve ter o valor *fem*.
Eliminado	[acc =]	Todos os valores do atributo *acc* são eliminados.

Tabela 2.1: Operações sobre os pares atributos-valor.

2.2.1 Léxico e pré-processamento

Léxico definido no XIP

O léxico pré-existente é aquele que provém da ferramenta de análise morfológica (e, eventualmente, do processo de anotação morfossintáctica), a que chamamos pré-processamento sintáctico. Para integrar este léxico no XIP é necessário definir o mapeamento entre as categorias e traços do pré-processamento sintáctico e aqueles que vão ser manipulados dentro do XIP.

O XIP permite a definição de entradas lexicais (ficheiros de léxico) e adicionar novas entradas que não foram guardadas no léxico pré-existente. Ter um vocabulário rico pode ser decisivo para uma melhoria na abrangência[5] do sistema. Deste modo, os dados lexicais necessários para a tarefa de RCEM foram acrescentados ao léxico geral utilizado pela análise sintática, sob a forma de léxico do XIP,.

No XIP, os ficheiros de léxico começam com a palavra `Vocabulary:`, desta forma indicam ao mecanismo que os ficheiros contêm léxico personalizado.

[5] *recall*

As regras lexicais têm como objectivo fornecer uma interpretação mais precisa dos *tokens* associados aos nós. Na Figura 2.7 está representada a sintaxe das regras lexicais (as partes entre parênteses são optionais).

```
lemma(: POS([features])) (+)= (POS)[features]
```

Figura 2.7: Sintaxe das regras lexicais.

Na Figura 2.8 está exemplificada uma regra lexical. Nesta nova entrada, `$US`, foram associadas novos traços booleanos: `meas` e `curr`. São estes traços que permitem definir esta entrada como uma unidade monetária.

```
$US = noun[meas=+, curr=+]
```

Figura 2.8: Exemplo de uma regra lexical.

Na Figura 2.9 está exemplificada a adição de um traço a uma entrada lexical. O traço `human:+` foi adicionado à entrada do nome `eleitor`, já existente no léxico.

```
eleitor: noun += [human=+].
```

Figura 2.9: Exemplo de uma entrada lexical onde foi adicionado um traço.

O processo de enriquecimento de entradas do léxico pré-existente com novos traços consiste na marcação de elementos linguísticos que funcionam como pistas contextuais, que servirão como auxiliares para a tarefa de identificação e classificação de EMs.

Tanto o processo de enriquecimento como o processo de adição de novas entradas foi realizada com base em listas de palavras já existentes ou criadas manualmente para o efeito. Por outras palavras, não são utilizados processos automáticos para o enriquecimento lexical.

2.2.1.1 Adaptação do pré-processamento

Além do enriquecimento do léxico, o XIP permite a definição de regras de desambiguação para as categorias gramaticais, que foram adaptadas para a tarefa de REM. A sintaxe geral para uma regra de desambiguação é a seguinte:

```
layer> readings_filter = |left_context| selected_readings |right_context|.
```

Figura 2.10: Sintaxe das regras de desambiguação.

As regras de desambiguação, tal como as regras de *chunking*, usam o conceito de camada e contexto. O lado esquerdo das regras de desambiguação contém `readings_filter` que especifica o conjunto de categorias e atributos que podem ser associadas a uma palavra. Por fim, o `selected_readings` da regra de desambiguação dá-nos a interpretação selecionada da palavra.

Existem quatro operadores principais usados nas regras de desambiguação:

- O operador < >: define os atributos específicos associados com a categoria;
- O operador [ ]: especifica o conjunto completo de características para uma categoria;
- O operador %: restringe a interpretação da palavra para uma solução;
- O operador < *: especifica que cada leitura tem de suportar as características listadas imediatamente depois.

Na Figura 2.11, a regra pode ser descrita da seguinte maneira: antes de uma forma verbal no infinitivo, não flexionado, a unidade lexical *pode* é a forma do verbo *poder*, e não a forma verbal do verbo *podar* (*e.g* O Pedro **pode** fazer isso.).

```
5> verb<lemma:podar>, verb<lemma:poder> = verb<lemma:poder> | verb[inf:+] | .
```

Figura 2.11: Exemplo de uma regra de desambiguação.

As regras de desambiguação auxiliam a tarefa de RCEM através de acréscimos lexicais. Por exemplo, a regra da Figura 2.12 permite desambiguar a palavra Natal (quadra festiva ou estado do Brasil):

```
20> noun[maj:+,surface:Natal] %= | noun[denot_time:+],
prep[lemma:de], art | noun[one_day=+,maj=+,proper=+].
```

Figura 2.12: Exemplo de uma regra de desambiguação com acréscimos lexicais.

Esta regra determina que depois de uma palavra relativa a um tempo, seguida da preposição *de* e, opcionalmente, um *artigo*, a palavra Natal corresponde à quadra festiva (*e.g.* Na semana do Natal).

2.2.2 Gramáticas locais para o REM

As gramáticais locais são ficheiros de texto que contêm regras de *chunking* e cada ficheiro pode conter regras de Depêndencia Imediata e Precedência Linear (Regras DI/PL)[6] e de sequência, estas regras serão definidas nesta Subsecção. São usados diferentes ficheiros de gramáticas locais essencialmente para capturar sequências de nós e para atribuir características relevantes. São usados diferentes ficheiros de acordo com as diferentes categorias de EMs. Por exemplo, enquanto o ficheiro `LGLocation` tem como objectivo guardar os nós relativos à categoria `LOCATION` (Localização), o ficheiro `LGPeople` tem como objectivo guardar os nós relativos ao tipo `INDIVIDUAL` e categoria `HUMAN` (Humano).

Depois do pré-processamento e da desambiguação, o XIP tenta corresponder as frases com as regras nos ficheiros de gramáticas locais. De forma a reconhecer as EMs é necessário analisar o texto em *chunking* e calcular as relações sintácticas entre potenciais EM e outros constituentes da frase [7]. As últimas fases do processamento das EM são, assim, o aproveitamento dos módulos de *chunking*, a construção de dependência e a propagação de traços. De seguida, serão descritas estas três fases.

[6] *Immediate dependency and linear precedence rules (ID/LP rules)*

Regras de *chunking*

A análise do texto em *chunks* é o processo no qual uma sequência de categorias são agrupadas numa estrutura. Este processo é feito através de regras de *chunking*, tais como:

- Regras DI/PL;
- Regras de Sequência.

Cada regra de *chunking* tem de ser definida na camada específica. Esta camada é representada por um número inteiro, compreendido entre 1 e 300. Na Figura 2.13 está representado um exemplo de duas regras em duas camadas diferentes.

```
1> NP = (art;?[dem]), ?[indef1]. //layer 1
2> NP = (art;?[dem]), ?[poss].  // layer 2
```

Figura 2.13: Duas regras em duas camadas diferentes

As camadas são processadas sequencialmente, da primeira à última, e cada uma pode conter apenas um tipo de regra de *chunking*.

As regras DI/PL são significamente diferentes das regras de sequência. Enquanto as regras DI descrevem os conjuntos não-ordenados de nós, as regras PL são usadas com as regras DI para estabelecer a ordem entre as categorias, por outro lado as regras de sequência descrevem as sequências ordenadas de nós. A sintaxe das regras DI é a seguinte:

```
layer> node-name -> list-of-lexical-nodes.
```

Figura 2.14: Sintaxe de um regra DI.

Considere o seguinte exemplo de uma regra DI:

```
1> NP -> det, noun, adj.
```

Figura 2.15: Exemplo de um regra DI.

Assumindo que *det*, *noun* e *adj* são categorias que já foram declaradas (consultar a Tabela A.1 no Apêndice A para a lista completa de categorias morfossintáticas), esta regra pode ser interpretada da seguinte maneira: *Sempre que existir uma sequência de determinante, nome e adjectivo, sem uma ordem específica, cria um chunk nominal*. Esta regra é aplicada a mais expressões do que aquelas desejadas, por exemplo todas as seguintes expressões são aceites: *o gato preto*, *gato o preto*, *gato preto o*, *o preto gato*, *preto o gato*, *preto gato o*. Para resolver este problema são usadas as regras PL. Ao serem associadas com as regras DI, elas podem ser aplicadas a uma camada particular ou serem tratadas como uma constante geral ao longo da gramática XIP. As regras PL têm a seguinte sintaxe:

```
layer> [set-of-features] < [set-of-features].
```

Figura 2.16: Sintaxe de um regra PL.

Considere o seguinte exemplo:

```
1> [det:+] < [noun:+].
2> [noun:+] < [adj:+].
```

Figura 2.17: Exemplo de um regra PL.

Segundo o exemplo da Figura 2.17, um determinante têm de preceder um nome , por sua vez um nome tem de preceder um adjectivo. Esta regra significa que a expressão *o preto gato* não será aceite. Por outro lado, *o gato preto* continuará válida.

É possível usar parênteses para expressar categorias opcionais e um asterisco para indicar que zero ou mais instâncias das categorias são aceites. A regra seguinte formula que o determinante é optional assim como ter tantos adjectivos quanto possível será aceite:

```
1> NP -> (det), noun, adj*.
```

Figura 2.18: Exemplo de uma regra DI usando *()* e $*$.

Tendo em conta as duas regras PL estabelecidas na Figura 2.17 e a regra DI da Figura 2.18, as seguintes expressões são aceites: *gato*, *gato preto*, *o gato preto*, *o gato preto bonito*.

Estas regras podem ser ainda mais restritas, através do uso do contexto. Por exemplo:

```
1> NP -> |det, ?*| noun, adj |?*, verb|.
```

Figura 2.19: Exemplo de uma frase restrita ao contexto através de uma regra DI.

A regra da Figura 2.19 define que um determinante tem de estar à esquerda de um conjunto de categorias e que um verbo tem de estar à direita para formar um *chunk*. Aplicando esta regra na frase: *O gato preto andou no telhado*, obtemos o seguinte *chunk*:

```
NP[gato preto]
```

Figura 2.20: Resultado das regras DI/PL numa frase.

Apesar de estas regras restringirem o padrão que formará o *chunk*, o contexto não fica guardado dentro do respectivo nó.

O outro tipo de regras de *chunking*, as regras de sequência, são conceptualmente diferentes porque descrevem uma sequência ordenada de nós, mesmo assim em termos de sintaxe são iguais às regras DI/PL. Contudo, existem algumas diferenças, tais como:

- As regras de sequência não usam o operador '->'. Por outro lado, usam o operador '=', no qual fazem correspondência com a sequência mais curta possível. De forma a corresponder a sequência mais longa possível é usado o operador '@=';
- Existe um operador para aplicar a negação(~) e outro para aplicar a disjunção (;);
- Ao contrário das regras DI/PL, o ponto de interrogação (?) pode ser usado para representar qualquer categoria no lado direito da regra;
- Regras de sequência podem usar variáveis.

A regra de sequência seguinte aceita as seguintes expressões: *Alguns rapazes*,*uns rapazes*, *nenhum rapaz*, *muitos rapazes* ou *cinco rapazes*:

```
1> NP @= ?[indef2];?[q3];num, (AP;adj;pastpart), noun.
```

Figura 2.21: Exemplo de uma regra de sequência.

O XIP oferece um formalismo que permite, entre outras coisas, exprimir regras de reescrita tomando em consideração, os contextos à esquerda e à direita da expressão regular a reescrever. As regras de gramáticas locais para as EM são usadas em dois tipos de situações:

- **Delimitação de EM complexas**

  Algumas das EM a reconhecer são constituídas por mais de uma palavra gráfica (unidade). Contudo, nas fases preliminares de pré-processamento, as várias unidades que constituem estas expressões apenas foram considerados individualmente, sendo função das gramáticas locais juntar esses elementos numa única EM. Considere como exemplo a regra da Figura 2.22.

```
1> noun[cargo=+,mwe=+,people=+] @= ?[cargo,maj],
(punct[hifen]), adj[lemma:"honorário", maj]; adj[lemma:mor].
```

Figura 2.22: Exemplo de uma regra de desambiguação com delimitação de EM complexas.

  Esta regra classifica um nó nominal complexo com dois traços (`cargo:+` e `people:+`), apenas se esse nó for uma sequência de palavras que começa com um elemento lexical que tem o traço `cargo:+`, seguido pelo adjectivo *honorário* ou *mor*. Assim, as sequências *Cônsul Honorário* ou *Sargento-mor* vão ser classificadas como nomes de cargo.

- **Utilização de contexto**

  Para algumas unidades lexicais, é o contexto imediato que permite reconhecer ou classificar uma EM. Considere o exemplo da Figura 2.23.

```
1> NOUN[org=+, institution=+] @= |?[lemma:governo,
maj:~],prep[lemma:de], (art)| ?[location].
```

Figura 2.23: Exemplo de uma regra de desmabiguação com utilização de contexto.

Esta regra classifica um nó complexo como organização institucional, se houver um contexto à esquerda constituído pelo nome *governo* seguido da preposição *de*, opcionalmente, seguido por um *artigo*, e um elemento lexical marcado com o traço `location`. Assim, uma expressão como o *governo de Lisboa* será classificado como uma organização institucional.

Estas regras podem ser aplicadas a sequências de palavras com categorias ambíguas. Relembramos que, a aplicação das regras locais nas EMs fazem-se depois da aplicação de um primeiro módulo de desambiguação e que grande parte das ambiguidades categoriais ainda não foram resolvidas. Estas gramáticas locais procedem, pois, a uma desambiguação suplementar, na medida em que, se houver emparelhamento com uma regra, serão selecionadas as categorias com que essas regras emparelharam.

Regras de dependência

A capacidade de extrair as dependências entre nós é importante porque pode fornecer um conhecimento mais rico e profundo do significado de um texto. As regras de dependência tomam a sequência dos nós identificados pelas regras de *chunking* e definem as relações entre elas. Esta secção apresenta uma visão global da sua sintaxe, os operadores envolvidos e alguns exemplos. Na sintaxe de uma regra de dependência, o parâmetro `if` e `<condition>` são opcionais. Uma regra de dependência apresenta a seguinte sintaxe:

```
|pattern| if <condition> <dependency_terms>.
```

Figura 2.24: A sintaxe de uma regra de dependência.

De forma a perceber o que é um `pattern` (padrão, em português), primeiro é fundamental perceber o que é uma Expressão Regular de Àrvore [7](TRE). A TRE é um modelo especial de expressões regulares usada no XIP de forma a estabelecer conexões entre nós distantes. Em particular, as TREs exploram a estrutura interna dos sub-nós através do uso do caracter chavetas ({}). O exemplo seguinte representa que a estrutura interna do nó NP têm de possuir um determinante e um nome:

```
NP{det, noun}.
```

Figura 2.25: Exemplo de uma TRE.

A TRE suporta o uso de vários operadores, tais como

- O ponto e vírgula (;), usado para indicar disjunção;
- O asterisco (*), usado para indicar "zero ou mais";

[7] *Tree Regular Expression (TRE)*

- O ponto de interrogação (?), usado para indicar "qualquer";
- O acento circunflexo (^), usado para explorar sub-nós para uma categoria.

Voltando às regras de dependência, o `pattern` contêm uma TRE que descreve as propriedades estruturais das partes do *input* da àrvore. A `condition` (condição, em português) é uma expressão booleana suportada pelo XIP (com a sintaxe adequada) e o `dependency_terms` (termos de dependência, em português) são as consequências da regra.

As primeiras regras de dependência a serem executadas são aquelas que estabelecem as cabeças dos *chunks* entre os nós, por exemplo a seguinte regra:

```
| NP#1{?*, #2[last]} |
    HEAD(#2, #1)
```

Figura 2.26: Exemplo de uma regra que identifica relações HEAD.

A regra identifica as relações HEAD (cabeça do *chunk*), por exemplo a frase *a bela rapariga* resulta na seguinte dependência: HEAD(rapariga, a bela rapariga).

Como já foi referido, o principal objectivo das regras de dependência é estabelecer relações entre os nós. Voltando à frase de exemplo, a Figura 2.27 representa o resultado de aplicar estas regras de dependência à frase: *O Sr.João foi à Índia*:

```
MAIN(foi)
HEAD(Sr.João,O Sr.João)
HEAD(Índia,a a Índia)
HEAD(foi,foi)
DETD(Sr.João,O)
DETD(Índia,a)
VDOMAIN(foi,foi)
MOD_POST(foi,Índia)
SUBJ_PRE(foi,Sr.João)
NE_PEOPLE_INDIVIDUAL(Sr.João)
NE_COUNTRY_ADMIN_AREA_LOCATION(Índia)
```

Figura 2.27: Exemplo das várias dependências dos nós numa frase.

As duas últimas regras da Figura 2.27 indicam que duas EMs forams identificadas e classificadas na frase: *Sr. João* foi classificado como `HUMAN INDIVIDUAL PERSON` (Humano Individual Pessoa, em português) e *Índia* foi classificada como `LOCATION CREATED COUNTRY` (Localização Criada País, em português). As etiquetas `NE_INDIVIDUAL_PERSON` e `NE_COUNTRY_ADMIN_AREA_LOCATION` são apenas para verificar que as EMs foram classificadas. O último ficheiro *XML* será criado de seguida, como último passo de todo o processo.

As outras dependências acima cobrem uma larga variedade de relações binárias como a relação entre:

- o núcleo dos *chunks* e os próprios *chunks* (`HEAD`);
- uma cabeça nominal e o determinante (`DETD`);
- a cabeça da Locução Prepositiva e a preposição (`SUBJ_PRE`).

A lista completa de todas as relações de dependência entre eventos está descrita na tese de mestrado realizada pela Viviana Cabrita [8].

Existem várias regras de dependência, considere o exemplo seguinte, que representa a classificação de EM:

```
| #1{?*, num[quant, sports_results]]} |
   if (~NE[quant,sports_results](#1))
    NE[quant=+,sports_results=+](#1)
```

Figura 2.28: Exemplo de uma regra de dependência destinada para a classificação de EMs.

Esta regra usa uma variável, representada por `#1`, que está localizada antes da primeira chaveta ({), logo está associada ao nó de topo. Esta regra afirma que, se após um nó existir um número com dois traços, quantidade(`quant`) e resultado de desporto(`sports_results`), e se esse nó ainda não foi classificado como EM com estas atributos, então o `#1` irá adiciona-lo ao nó de topo, de forma a classificá-lo como um resultado de desportoAMOUNT SPORTS_RESULTS. O nó de topo é que classifica, visto que a variável está associada a ele. Se a variável estivesse antes do nó `num`, nesse caso apenas o sub-nó seria classificado.

Note-se que o uso do operador de negação (~) dentro da declaração condicional. A sintaxe do XIP também permite o uso do operador '`&`' para a conjunçao e o operador '`|`' para a disjunção nas declarações condicionais. Os parênteses também são usados para agrupar declarações e estabelecer prioridades, como na maioria das linguagens de programação.

Propagação de traços

O único módulo que falta analisar é da propagação de traços e é dele que falaremos de seguida. Um dos problemas com que se defronta a tarefa de RCEM consiste na resolução de casos de metonímia. A fim de capturar este fenómeno, é necessário ter em consideração um contexto relativamente alargado. Logo, as regras contextuais das gramáticas locais não têm o formalismo mais adequado para representar esse tipo de contexto. A frase (2.1) exemplifica este problema.

O navio Sagres explorou o oceano. (2.1)

O nome *navio Sagres* está marcado no léxico como um nome de um barco. No fim da cadeia de processamento, esta unidade lexical seria classificada como uma EM do tipo `LOCAL`. No entanto, neste contexto sintáctico, como sujeito de um verbo como *explorou*, o *navio Sagres* não designa o espaço fisico mas sim os trabalhadores da organização, eventualmente, um grupo de marinheiros. É possível corrigir a classificação deste nome graças às dependências previamente calculadas, nomeadamente a relação de sujeito entre o nome e o verbo, o nome passa a ser tratado, então, como `ORGANIZACAO`.

Esta correção ao nível das dependências resulta na correta classificação do nome *navio Sagres*, sempre que este é sujeito ou agente da passiva de um verbo como *explorar*.

O exemplo seguinte mostra uma regra que permite transformar uma EM de tipo geográfico em EM de tipo organização quando ela é sujeita ou agente da passiva do verbo *explorar*.

```
if ( ^NE[local:+,admin_area:+](#1) &
            ( SUBJ(?[lemma:explorar],#1)
            | AGENT(?[lemma:explorar],#1)
            )
   )
NE[features=\~,org=+,administration=+](#1)
```

Figura 2.29: Exemplo de uma regra que altera o tipo de uma EM.

A propagação é um mecanismo que permite conservar a informação sobre EM previamente calculadas e propagar essa informação ao resto da análise do texto. Este processo parte do pressuposto de que, num mesmo documento, novas EM são introduzidas num contexto suficientemente rico para que possam ser classificadas de forma não ambígua; no entanto, muitas vezes, essas EM são retomadas nesse mesmo texto mas noutro contexto. Trata-se, tipicamente, do caso de nomes de pessoas mas pode também acontecer com outro tipo de entidades.

O XIP oferece, além das operações habituais para o processamento linguístico, uma Linguagem Dedicada [8], que pode ser usada para algumas tarefas simples, tais como contagens de ocorrências ou de relações.

A propagação é um mecanismo poderoso, que permite aumentar a *recall* de um sistema de RCEM. No entanto, pode também ter efeitos perversos, sobretudo se a entidade inicial não tiverem sido correctamente classificada.

Para concluir este capítulo, a integração do módulo de RCEM na cadeia de processamento é motivada por vários factores: em particular, o reconhecimento das EM permite melhorar os resultados dos outros módulos de processamento linguístico.

O facto de se ter acesso à estrutura sintáctica permite ir mais longe na tarefa de RCEM [7]. Nesta tarefa, é possível, assim, fazer *chunking* e calcular as relações sintácticas e as dependências entre potenciais entidades mencionadas e outros constituintes da frase [7]. As últimas fases do processamento das EM consistem, assim, no aproveitamento dos módulos de *chunking*, de construção de dependências e de propagação de traços. Todos estes módulos se encontram integrados no XIP.

[8] *Scripting Language*

Capítulo 3

Corpus

Neste Capítulo é descrito o domínio textual, assim como a estrutura dos documentos. Também é descrito o processo de tratamento dos documentos, desde o formato fornecido (*.pdf*) até a um formato que a STRING consiga processar.

3.1 Domínio

O domínio textual usado neste projeto é a correspondência da Marinha Portuguesa, um ramo das Forças Armadas Portuguesa. Esta instituição é dotada de autonomia administrativa e integra-se na administração do Estado através do Ministério da Defesa Nacional.

3.1.1 Estrutura da Marinha

A organização da Marinha rege-se pelos princípios da eficácia na perspetiva da missão, da flexibilidade ao nível operacional, da otimização estrutural e do equilíbrio genérico, em articulação, e em alguns casos complementaridade, com o Ministério da Defesa Nacional, o Estado-Maior General das Forças Armadas e com os outros ramos das Forças Armadas.

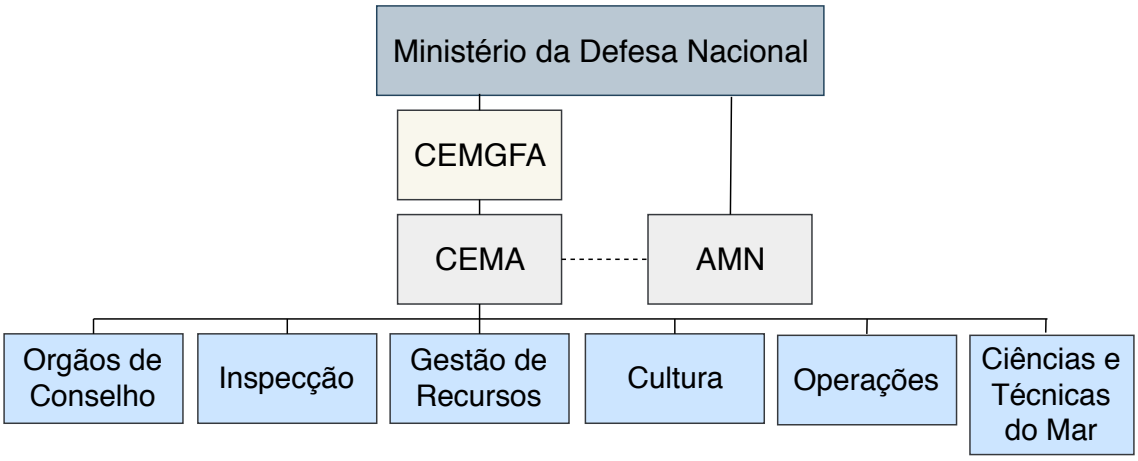

Figura 3.1: Organização estrutural da Marinha Portuguesa. [17]

O Chefe do Estado-Maior da Armada (CEMA) depende do Chefe do Estado-Maior General das Forças Armadas (CEMGFA) para efeitos operacionais, e do Ministro da Defesa Nacional para a administração de recursos. O CEMA é por inerência a Autoridade Marítima Nacional, como está representado na Figura 3.1.

Os órgãos de apoio direto ao Almirante CEMA concentram a responsabilidade pela formulação estratégica e planeamento no que concerne a toda a visão estratégica para a Marinha. Como está representado na Figura 3.1, o CEMA é responsável por seis unidades: Orgãos de Conselho, Inspecção, Gestão de Recursos, Cultura, Operações e Ciências e Técnicas do mar. A Gestão de Recursos têm quatro superintendências (do Material, do Pessoal, das Finanças e das Tecnologias da Informação), que cabe administrar os respetivos recursos. O comando da componente naval, com os respetivos elementos, responde pelo comando e emprego dos meios e recursos atribuídos. Os órgão responsáveis pelas ciências e técnicas do mar são o Instituto Hidrográfico e a Escola Naval. A Academia de Marinha e a Comissão Cultural de Marinha são os órgãos de natureza cultural. A atividade inspetiva na Marinha é garantida pela Inspeção-Geral da Marinha [17].

Os órgãos centrais de administração e direção (também conhecidos como Gestão de Recursos) têm caráter funcional e visam assegurar a direção e execução de atividades específicas essenciais, tais como a gestão de recursos humanos, materiais, financeiros e de informação, como está ilustrado na Figura 3.2. Cada área é designada de *superintendência* e tem um objectivo distinto, como descrito abaixo.

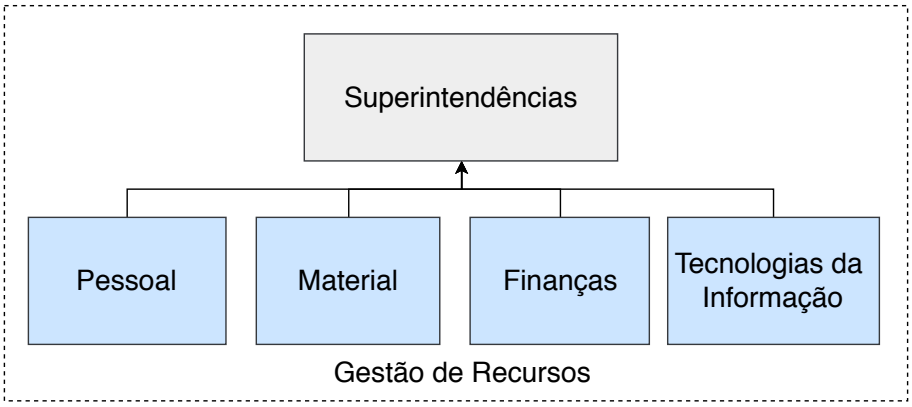

Figura 3.2: Estrutura da Gestão de Recursos da Marinha Portuguesa.[2]

A Gestão de Recursos da Marinha divide-se em quatro superintendências distintas, como está representado na Figura 3.2:

- **Pessoal:** administração dos recursos humanos, formação e saúde;
- **Material:** administração dos recursos materiais;
- **Finanças:** administração dos recursos financeiros;
- **Tecnologias da Informação:** administração dos recursos informacionais.

Todas as superintendências atuam de forma a não prejudicar a competência específica de outras entidades.

O domínio textual específico deste projecto é a Superintendência das Tecnologias da Informação (STI), uma vez que todo o *Corpus* provém de documentos que circularam nesta superintendência.

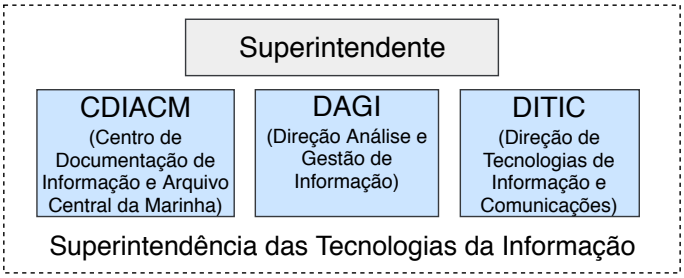

Figura 3.3: Componentes da Superintendência das Tecnologias da Informação (STI). [1]

A Superintendência das Tecnologias da Informação têm quatro entidades, como está representado na Figura 3.3:

- **Superintendente das Tecnologias da Informação:** dispõe de autoridade funcional e técnica sobre todos os órgãos da Marinha no domínio dos recursos informacionais, compreendendo a análise, gestão e arquivo da informação, e os sistemas, infraestruturas de suporte e tecnologias da informação e de comunicações, sem prejuízo da autoridade funcional do Superintendente do Material no âmbito das unidades navais;
- **CDIACM:** assegura o exercício da autoridade técnica no âmbito da arquivística e documentação;
- **DAGI:** assegura o exercício da autoridade técnica no domínio da gestão e análise da informação, da arquitectura de referência, administração de dados, estatística e investigação operacional;
- **DITIC**: assegura o exercício da autoridade técnica no domínio das comunicações e sistemas de informação (CSI) e tecnologias de informação e comunicações (TIC) da Marinha, sem prejuízo da competência específica de outras entidades no mesmo âmbito.

Com uma abordagem do geral para o específico foi realizada uma descrição do domínio, partindo do Ministério da Defesa Nacional, passando pelo Chefe do Estado-Maior da Armada e todos os orgãos de apoio direto, terminando na Gestão de Recursos. De todos os orgãos de apoio existe um que tem especial relevância para este projecto, os orgãos centrais de administração e direção. Como este projeto é na aréa da Ciência da Computação, o domínio textual abordado concentrar-se-à na administração dos recursos informacionais, que por sua vez é designada por Superintendência das Tecnologias da Informação (STI).

3.2 Documentos

O *Corpus* foi extraído de documentos recebidos e enviados das quatro unidades da STI, representadas na Figura 3.3. Os documentos estão organizados por ano, de 2015 a 2019[1]; dentro do ano divide-se em três grupos: Entradas, Internos e Saídas; por fim, dentro da origem dos documentos estão os quatro departamento da STI. O nome dos documentos é formado por três elementos separados por um hífen (-), o primeiro é a especificação da origem (*E* se o documento foi entregue naquela unidade, *S* no caso de ter saído); o segundo elemento é um identificador único, um número incremental; o terceiro são as siglas da unidade onde o documento está situado. De forma a a atingir a coerência entre os documentos foi removido a especificação dos meses no ano 2014 e 2015, também foi removido as pastas em que a origem dos documentos é Interna. Estas remoções são justificada pelo facto de apenas o ano de

[1]Apenas no ano de 2014 e 2015 existe a divisão por meses após a divisão por departamentos.

2014 e 2015 ter a especificação por meses e a pasta de documentos em que a origem é Interna nem sempre está presente.

O *Corpus* possui um total de 7302 documentos. Os documentos são maioritariamente das unidades STI e DITIC. Esta predominância tem vindo a aumentar ao longo dos anos, como ilustra a Figura 3.4. Quanto ao seu conteúdo, os documentos são muito variados, sendo que 64,89% dos documentos têm a mesma estrutura, designados por normalizados. Os restantes documentos são não normalizados, podendo ser faturas, recibos, faxes, memorandos, diplomas ou *e-mails* do Gabinete do Chefe do Estado-Maior da Armada.

Este *Corpus* vai ser usado para aprendizagem automática em dois projetos com classificadores distintos. Um projeto usará o número de processo (Subsecção 3.2.3), presente apenas nos documentos normalizados, porém, o outro projeto usará a tabela de distribuição(Subsecção 3.2.6), presente em 90,85% dos documentos. Em suma, o *Corpus* é constituído por diferentes tipos documentais em que todos têm uma tabela de distribuição, contudo, cerca de 65% dos documentos têm um número de processo.

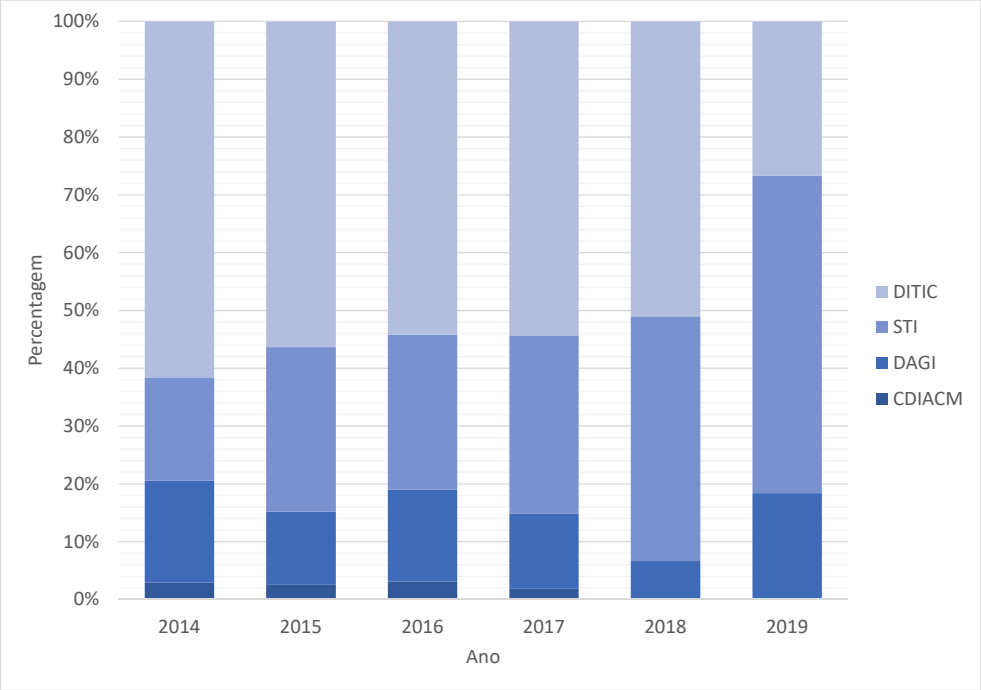

Figura 3.4: Distribuição dos documentos pelas unidades ao longo dos anos.

3.2.1 Estrutura dos documentos

Os documentos normalizados têm uma estrutura pré-definida, logo são os únicos documentos que requerem uma análise descritiva. A estrutura está organizada em seis segmentos: *cabeçalho*, *número do processo e do documento*, *assunto*, *referência(s)*, *destinatário(s)*, *corpo* e *assinatura*. Existe um elemento optional, o *anexo*. De seguida, será analisado, individualmente, cada um dos segmentos.

3.2.2 Cabeçalho

Figura 3.5: Cabeçalho dos principais documentos da Marinha.

Na Figura 3.5, as letras a preto identificam a instituição na qual o documento pertence, esta informação está presente em todos os documentos normalizados. Assim como, as letras a vermelho identificam a origem do documento, podendo indicar a Unidade, Serviço, Estabelecimento ou Organismo. Por fim, também é indicado o Local e a Data do documento.

3.2.3 Nº do Documento e Processo

Estes dois números estão especificados logo abaixo do cabeçalho. O número do documento é um identificador do documento na sua unidade de origem. O número do processo é um classificador do documento. Este classificador tem 4 níveis separados por pontos, cada nível é representado por um número inteiro, sendo o último opcional. O primeiro número tem 3 unidades e é múltiplo de 10, este número representa a função, a atividade da Marinha na qual o documento está associado; por sua vez, o segundo número tem 2 unidades, como os próximos, é múltiplo de 5 e representa a Subfunção; de seguida, o terceiro representa a Série, ou seja, uma atividade específica da Marinha; por fim, o quarto número representa a Subsérie. Existem 17 funções, 118 subfunções e 1.181 números de processos diferentes.

Por exemplo, o documento com o número de processo 080.10.07 está associado à função "Gestão de Recursos Humanos", a subfunção "Avaliação de desempenho e de mérito"e a série "Provas de aptidão física (PAF)". Portanto, este documento é relativo a provas de aptidão física obrigatórias, realizadas pelos militares, nomeadamente convocatórias do Centro de Educação Física da Armada (CEFA), agendamento e resultados das provas de aptidão (apto/não apto).

3.2.4 Assunto

O título do assunto está escrito em letras maiúsculas sublinhadas alinhadas à esquerda e não *justificadas*. Esta informação serve para o leitor identificar o tema do documento.

3.2.5 Referência(s)

As referências são colocadas pela ordem em que são mencionadas no texto. Se forem duas ou mais, utilizam-se letras para as identificar, seguidas do sinal de fechar parêntesis; *e.g* "a)", "b)", etc. Quando aplicável, é mencionado,

entre parêntesis, o anexo onde constam. Esta informação identifica os documentos que estão associados a este documento.

3.2.6 Destinatários

Após as referências, existe uma secção onde é especificado quem são os destinatários do documento. Os destinatários podem ser de dois tipos, os que devem agir com base no documento ou os que apenas devem tomar conhecimento. Os primeiros são indicados após a contração *à* ou *ao*. Os segundos são indicados após a frase sublinhada: "*Para conhecimento:*"e a mesma contração do anterior. Esta informação irá ajudar a redireccionar os documentos para as unidades corretas.

Os destinatários são especificados na Tabela de Distribuição, representada na figura 3.6, situada no final da primeira página do documento. Nesta tabela é adicionado a letra "C"nas unidades que deve tomar conhecimento do documento e a letra "A"naquelas que devem tomar acção sobre o documento. A tabela não é preenchida apenas com a correspondência direta entre os destinatários especificados, pois, existe dependência entre as unidades e os documentos. Por exemplo, sempre que a STI envia um documento para a Inspecção-Geral da Marinha, todas as componentes desta superintendência (CDIACM,DAGI, DITIC) deverão tomar conhecimento, mesmo quando não identificadas.

Na Figura 3.6 estão identificadas algumas unidades pelas suas siglas, parte das unidades foram enunciadas na secção 3.1.1 enquanto as restantes têm os seguintes significados:

- **C/GAB:** Gabinete do Superintendente das Tecnologias da Informação;
- **SERV.PART:** Serviço Particular;
- **ADJ.SEC1:** Entidade Contabilística.

As seguintes unidades também podem aparecer na tabela de distribuição, contudo não foram contabilizadas no *Corpus* devido ao reduzido número de ocorrência:

- **DAP:** Departamento de Apoio;
- **PMO:** Escritório de Gestão de Projetos;
- **DSUP:** Depósito de Suprimentos.

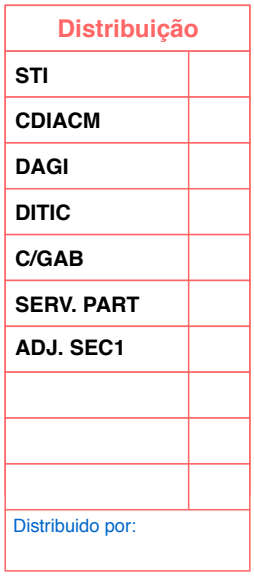

Figura 3.6: Tabela de Distribuição por preencher.

3.2.7 Corpo

O corpo é o texto que contém a informação que se pretende comunicar. Esta informação pode ser um *parecer*, quando se destina a expor uma opinião sobre problemas; uma *informação*, quando se destina a elucidar uma entidade ou serviço relativamente a um determinado assunto; por fim, pode ser uma *proposta*, quando se destina a sugerir procedimentos inovadores ou novas disposições legais, ou que envolve, concretamente, aspectos administrativos de que resultam encargos financeiros.

3.2.8 Assinatura

Este tipo de documento tem sempre uma assinatura. A assinatura é sempre seguida da frase "*Com os melhores cumprimentos*", além disso, a assinatura é acompanhada com a identificação do autor, nome completo e patente.

3.2.9 Anexo

Por fim, o anexo, tal como o nome indica, é um ou mais documentos que foram adicionados ao principal. Estes documentos podem ser normalizados ou não e são adicionados após o documento principal. Apesar do anexo ser opcional está presente em 74% dos documentos.

3.3 Tratamento dos documentos

O tratamento dos documentos divide-se em quatro fases: conversão, filtragem, remoção e segmentação. Esta cadeia de tratamento tem como objectivo corrigir erros de conversão, limpar informação desnecessária e segmentar a informação por secções. Desta forma, os dados presentes no *Corpus* melhoraram a sua qualidade ficando mais perto da informação presente nos documentos fornecidos. Assim, o formato final dos documentos é o formato de

texto, desta forma a STRING consegue processá-los. Este tratamento foi realizado através programas desenvolvidos em *Python 3.7.1* com a biblioteca de expressões regulares (*Lib/re.py*). Na Figura 3.7 estão ilustrados as várias fases dos documentos, assim como as tarefas correspondentes.

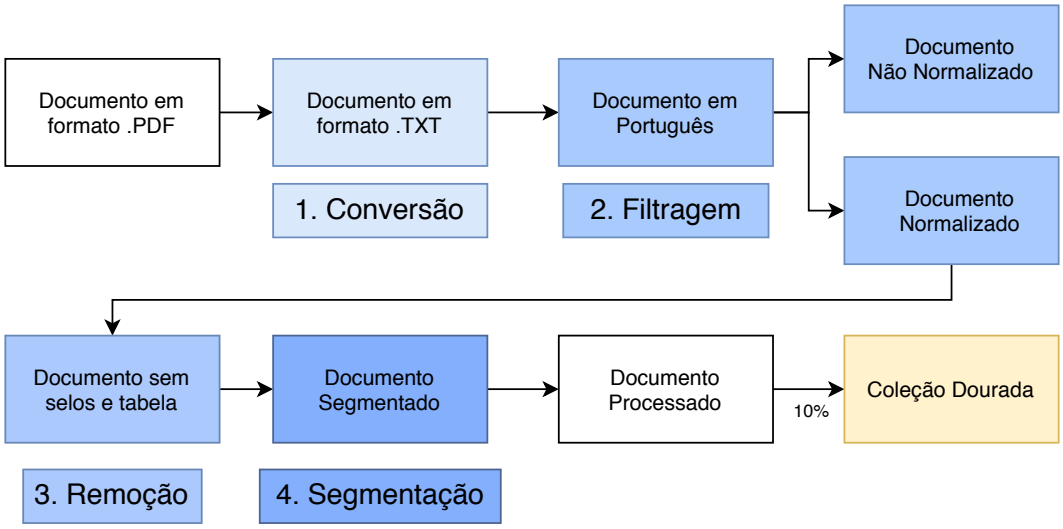

Figura 3.7: Evolução dos documento ao longo do processo de tratamento.

Durante este processo a estrutura das pastas e o nome dos documentos foi alterada. No nome dos documentos está presente a origem e a unidade, logo, essas duas pastas foram eliminadas, ficando apenas a pasta do ano dos documentos. Relativamente ao nome dos documentos, foram alterados de forma a atingir a especificação anteriormente descrita no Secção 3.2. Nesta secção será descrito o processo pelo qual os documentos foram submetidos.

3.3.1 Conversão

Os documentos fornecidos pela Marinha Portuguesa estavam, inicialmente, no formato *Portable Document Format* (*pdf*). Como a ferramenta de PLN usada suporta apenas textos no formato texto (*txt*), um dos passos do tratamento do *Corpus* passou por converter estes documentos de *pdf* para *txt*. Para este processo foi utilizada uma ferramenta de Reconhecimento Óptico de Caracteres[2], ABBYY fine reader 12. Este OCR tem características relevantes para o correto reconhecimento do *Corpus*, tais como:

- processar documentos em Português com auxílio de um dicionário da língua portuguesa;
- preservar a estrutura do documento ao longo das páginas;
- reter todos elementos do documento: cabeçalho e rodapé, diagramas, gráficos e tabelas;
- reconhecer assinaturas.

A STRING apenas suporta textos em português, logo qualquer documento em inglês foi ignorado. O reconhecimento dos documentos em inglês foi realizado através da procura das 100 palavras mais comuns na língua inglesa nos documentos. Destas palavras existem três que existem na língua portuguesa, as palavras: *a*, *as* e *use*. De forma a melhorar a procura de textos em inglês, estas palavras foram removidas da nossa lista de 100 palavras. Um documento foi considerado escrito em inglês quando mais de 20% das palavras dos documentos pertenciam à

[2]Em inglês *Optical Character Recognition (OCR)*

lista [21]. No final da fase de conversão foram removidas todas as linhas com espaços em branco. Esta medida foi executada para facilitar o processamento do texto nas fases seguintes.

3.3.2 Filtragem

Os tipos de documentos são variados, logo, também foi necessário separar os documentos normalizados dos não normalizados de forma a realizar um tratamento específico para cada caso. Os primeiros documentos filtrados foram aqueles descritos na secção 3.2, pois, são os documentos em maior quantidade. Estes documentos normalizados têm todos a mesma estrutura e podem ser de três tipos: propostas, notas ou ofícios.

De forma a identificar um texto como normalizado foi utilizada a secção cabeçalho, descrita na Subsecção 3.2.2, uma vez que esta secção é única para este tipo de documentos. Foi definido um padrão para cabeçalho através de uma expressão regular para classificar o documento como normalizado.

A distribuição destes documentos ao longo dos anos está apresentada na Figura 3.8. Existe uma diminuição da percentagem de documentos normalizados do ano 2014 para o ano 2016, contudo a percentagem sobe do ano 2016 até 2019. A distribuição dos documentos normalizados dentro dos departamentos é do tipo parábola, ou seja, diminui gradualmente entre 2014 e 2016 e aumentou gradualmente entre 2017 e 2019. Com esta análise, prevemos que a percentagem de documentos normalizados no ano de 2020 será superior a 90%, relembramos que a média deste documentos nos últimos 6 anos é de 74%.

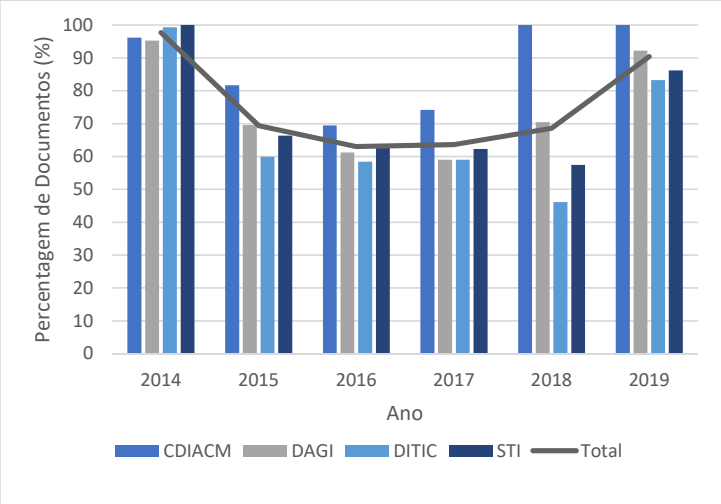

Figura 3.8: Distribuição dos documentos normalizados nos departamentos ao longo dos anos.

3.3.3 Remoção

Os documentos normalizados possuem informação desnecessária para o RCEM, logo foi removida. Os seguintes elementos foram identificados como desnecessários:

- **Selos de Entradas e Saída:** Na Figura 3.9 está exemplificado o selo de entrada e de saída. Nestes selos são registadas informações acerca da entrada ou saída dos documentos nas unidades. Eles possuem o número do

documento, o número do processo, a unidade, a data e, opcionalmente, quem processou o selo. Estes selos são adicionados ao documento de forma aleatória sem ocultar o texto do documento.

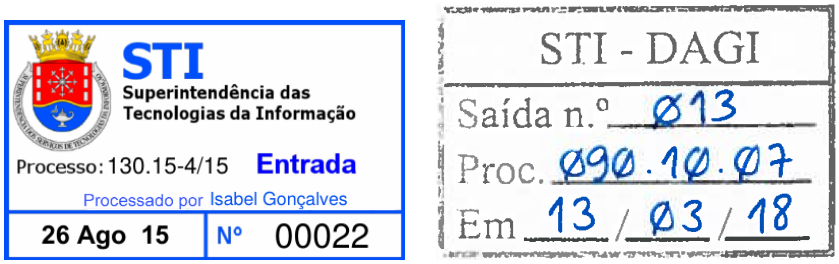

Figura 3.9: Exemplo de um selo de entrada (esquerda) e de saída (direita).

- **Selo dos destinatários:** Na Figura 3.10 é ilustrado um exemplo do selo dos destinatários. Este selo é proveniente do gabinete do CEMA ou do AMN e é referido o número do documento, a data, o destinatário e quem deve tomar conhecimento. Também pode ser especificado o prazo de entrega e podem conter outras observações. Este selo está situado normalmente junto ao cabeçalho.

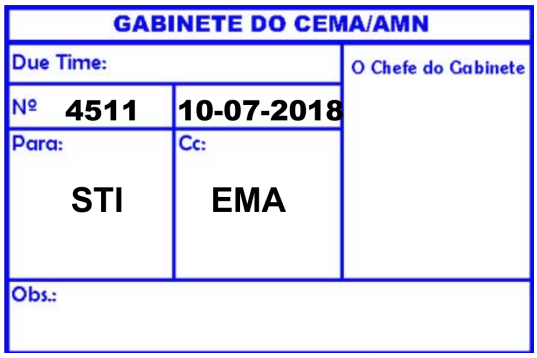

Figura 3.10: Exemplo de um selo para redirecionar o documento.

- **Tabela de Distribuição:** Na Figura 3.6 está ilustrada a tabela de distribuição. Para este projeto, esta informação é considerada desnecessária, contudo ela é importante para o projeto seguinte, pois será usado numa tarefa de aprendizagem automática;
- **Escrita à mão:** Ao longo do documento existem vários elementos escritos à mão, não só nas assinaturas, mas também para acrescentar informação ao documento. Esta escrita, por vezes, sobrepõe-se à informação presente no documento criando erros na conversão do texto.

Como foi referido anteriormente, as tabelas de distribuição serão necessárias para a fase seguinte do projeto, logo, antes de eliminar este elemento, a sua informação foi guardada num documento em formato *JavaScript Object Notation (JSON)*, como identificador do documento foi concatenado o ano e o nome do documento. Para cada documento foi associada uma lista de 7 elementos, descritos na Subsecção 3.2.6. Esta lista tem a ordem apresentada na Figura 3.6 e cada elemento pode ter 3 valores: "0" quando não há nenhuma ação associada a essa unidade, "A" quando essa unidade deve tomar ação sobre o documento ou "C" quando essa unidade deve tomar conhecimento acerca do documento. Na Figura 3.11 está um exemplo de uma entrada do documento *JSON*.

```
"2017/E-033-DAGI": [
 "0",
 "0",
 "C",
 "A",
 "C",
 "0",
 "0"
]
```


Figura 3.11: Exemplo da representação da tabela de distribuição para um documento.

Esta entrada representa que para o documento "E-033-DAGI"de 2017 a unidade DAGI e o Gabinete do Superintendente das Tecnologias da Informação (C/GAB) devem tomar conhecimento do documento, além disso, a unidade DITIC deve tomar ação sobre o documento.

As tabelas de distribuição foram identificadas em 90,85% dos documentos. Os documentos que não possuem a tabela de distribuição foram removidos do *Corpus*, visto que esta informação será a classificação do documento e sem ela não será possível realizar a aprendizagem automática no próximo projecto. Contudo, 458 documentos (6,2%) foram identificados com a tabela de distribuição mas sem tradução direta para o documento *JSON*, devido a erros de leitura por parte do *OCR*. Para estes casos foi necessário analisar os documentos em formato *pdf* e escrever, manualmente, a informação no documento *JSON*.

A Figura 3.12 representa a distribuição dos selos nas várias unidades ao longo do tempo. O número de selos de cada documento normalizado decresce ao longo do tempo, contudo a distribuição dos selos nas várias unidades por ano mantém-se, relativamente, constante.

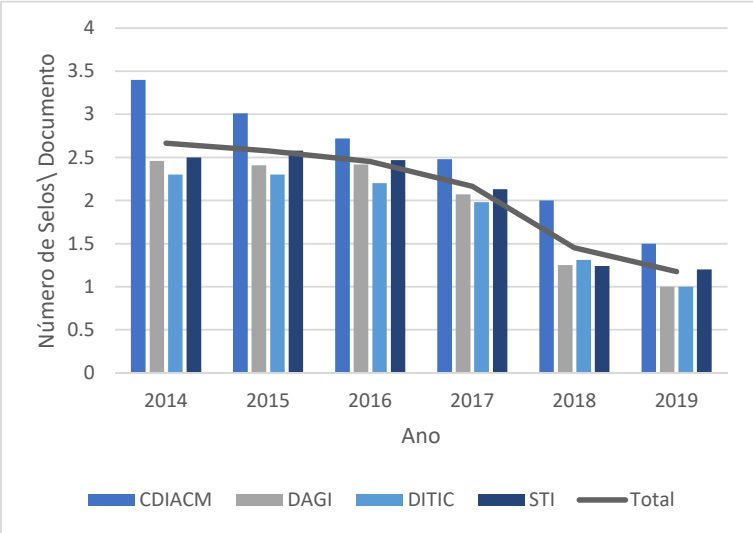

Figura 3.12: Distribuição dos selos dos documentos principais nos departamentos ao longo do tempo.

Nesta fase também foram eliminadas e corrigidas palavras devido a erros do *OCR*. Para o processo de eliminação, foram identificadas as palavras que ocorreram uma vez, sendo cerca de 53% das palavras. Visto que cerca 31% das palavras são números, apenas as palavras com menos de 50% de caracteres numéricos foram adicionadas

à lista para eliminar.

Estas palavras foram submetidas a dois filtros para identificar se a palavra deveria ser eliminada ou corrigida. Os dois filtros usados para identificar as palavras a eliminar foram os seguintes:

- Menos de 15% de vogais na palavra. Por exemplo, a palavra *ÍMLSHftRRft* é claramente um erro do *OCR*;
- Mais de 15% de caracteres especiais, que não sejam letras ou números, na palavra. Por exemplo, a palavra *Tif1$rr_!rfo* tem cerca de 30% de caracteres especiais,logo foi removida do *Corpus*.

Depois desta filtragem, 52.675 palavras, cerca de 19,3% das palavras que ocorreram uma vez, foram identificadas como palavras a eliminar. Após a eliminação das palavras, o total de palavras diferentes do *Corpus* diminui cerca de 10%.

Para realizar a correcção das palavras foi usada esta lista de palavras ordenada por ordem alfabética, pois, normalmente, as palavras incorrectas estão alfabeticamente próximas das palavras corretas. Desta forma, o processo de reconhecimento de palavras a corrigir foi agilizado. Foram identificadas 15662 palavras incorrectas para 2160 palavras diferentes, cerca de 7 palavras incorrecta por palavra. As palavras incorrectas ocorrem várias vezes ao longo do *Corpus*, foram corrigidas 131.828 palavras, cerca de 2,28 % do número total de palavras. Apesar da ferramenta *OCR* ser para documentos em português, cerca de 46% das corrigidas têm c de cedilha ou acentos. Logo, concluímos que o *OCR* tem dificuldades na leitura de documentos em português.

A remoção dos elementos desnecessários reduziu, em média, 58 palavras (4,8%) por documento. Desta forma, foi reduzido a complexidade e melhorada a verosimilhança dos dados nos documentos.

3.3.4 Segmentação

A segmentação consiste em dividir os dados pelos seus segmentos através de etiquetas, aumentando o conhecimento da estrutura interna do documento. Esta divisão permitirá pesar os dados dos vários segmentos de forma diferente. Por exemplo, na tarefa de aprendizagem automática, pode-se considerar o texto do título mais relevante do que o texto do corpo.

Os segmentos foram identificados através de padrões usando expressões regulares. De forma a validar os diferentes segmentos no documentos, estes foram convertidos em formato *eXtensible Markup Language (XML)*. Assim, cada segmento foi identificado com uma etiqueta específica de abertura e fecho, próprio desta linguagem de marcação.

Para validar os diferentes segmentos foi desenvolvido um ficheiro *XML Schema Definition (XSD)*. Com o *XSD* foi possível certificar que todos os documentos normalizados preenchiam os requisitos. Definimos que para ser considerado um documento normalizado tinha, obrigatoriamente, de ter:

- Cabeçalho, identificador do documento normalizado;
- Tabela de Distribuição, classificador de aprendizagem automática;
- Número de Processo, classificador de aprendizagem automática;
- Título, indica o assunto do documento;

- Corpo, indica o texto principal do documento.

Os restantes segmentos: número de documento, referência, destinatário, assinatura e anexo; são opcionais. Contudo, como referido na Subsecção 3.2.9, 3 em cada 4 documentos têm anexo, logo é um segmento relevante para este tipo documental. Por vezes, o corpo contém apenas uma referência ao anexo, tornando este último a principal fonte de informação do documento.

Durante esta fase surgiram dificuldades na identificação dos segmentos devido a erros do *OCR*, tornando a estrutura dos documentos irregular. Os erros surgiram devido aos selos e tabelas, que dificultaram a leitura correta. Para além da informação desnecessária referida na secção 3.3.3, existe outro tipo de informação que dificultou a detecção dos segmentos, como é o caso das imagens, do cabeçalho e do rodapé. Apesar desta informação desnecessária ter sido removida, a leitura aconteceu primeiro, influenciando a estrutura do documento de texto. De forma a resolver os erros foi necessário identificar, manualmente, alguns segmentos. Esta identificação foi realizada através da consulta dos documentos fornecidos. Foram identificados 406 documentos (5,5%) sem título; 448 (6,1%) sem corpo; e, finalmente, 228 (3,1%) documentos sem número de processo.

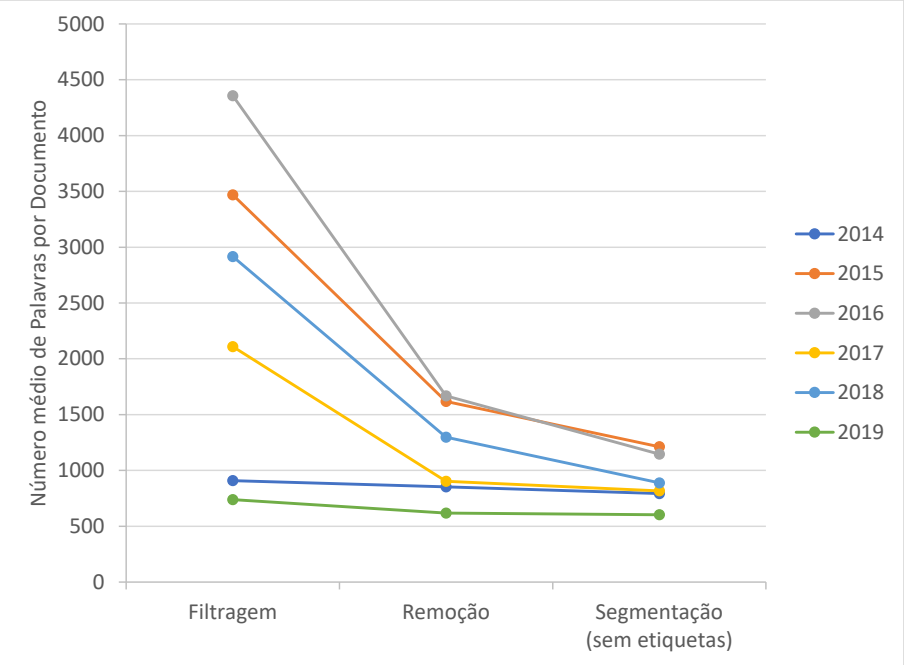

Figura 3.13: Número médio de palavras nos documentos ao longo do processo.

Na Figura 3.14 está representado a evolução do número de palavras ao longo do processo de tratamento. O número de palavras dos documentos normalizados vai aproximando entre os quatro departamentos, logo, concluímos que a complexidade dos dados dos documentos foi reduzindo ao longo do processo.

3.4 Coleção Dourada

A Coleção Dourada é uma coleção de documentos anotados que permite avaliar a desempenho da cadeia de PLN. Esta anotação é realizada manualmente, para tal foram escolhido, aleatoriamente, 210 documentos, cerca de

3% do *Corpus*.

Devido à variação da distribuição dos documentos nas diferentes unidades ao longo do tempo (Figura 3.4), assim como a redução do número de selos (Figura 3.12), concluímos que existe uma evolução significativa dos documentos ao longo do tempo. De forma a ter uma coleção mais representativa da realidade, valorizámos os documentos mais recentes. Para tal, distribuímos, equitativamente, os documentos ao longo dos anos, usando a fórmula dos números triangulares (3.1). Esta fórmula diz-nos por quanto é que temos de dividir a Coleção Dourada para atingir uma distribuição proporcionalmente crescente.

$$T_n = \sum_{k=1}^{n} k = 1+2+3+\ldots+(n-2)+(n-1)+n = \frac{n(n+1)}{2} \tag{3.1}$$

Por exemplo, para o nosso caso, sabendo que o intervalo de anos é entre 2014 e 2019, logo n é igual a 6, substituindo na equação obtemos o número triangular 21, ou seja, temos de dividir o número total de documentos da Coleção Dourada por 21 que dá cerca de 10 documentos (4.76%). Deste modo, a nossa Coleção Dourada tem 10 documentos no ano de 2014; 20 documentos, 9.52% (2 x 4.76%), em 2015; e assim sucessivamente até atingir 60 documentos, 28.57% (6 x 4.76%) em 2019. Dentro da pasta relativa ao ano, a distribuição dos documentos foi uniforme, dado que não existe nenhuma variação relevante entre as unidades dos documentos.

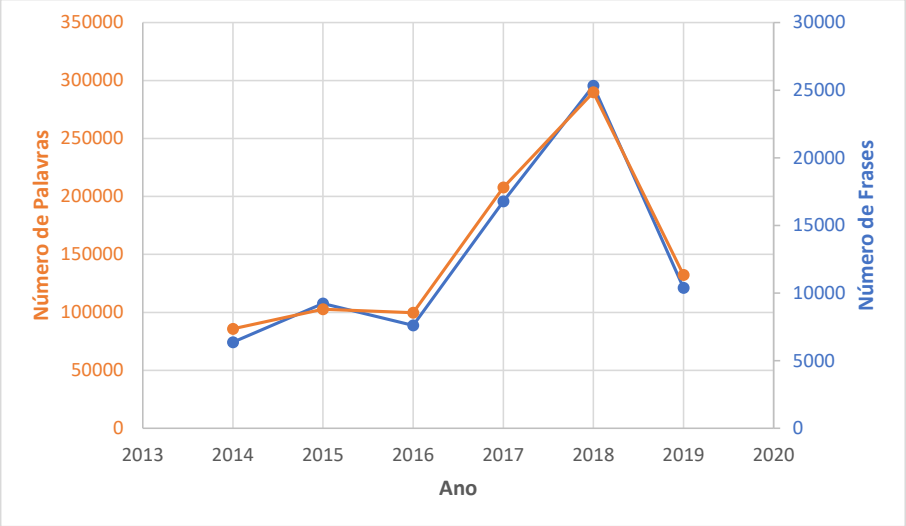

Figura 3.14: Número palavras e frases nos documentos ao longo dos anos.

Como já foi referido, o aumento do número de documentos ao longo dos anos é constante, contudo o número de palavras e frases não tem o mesmo comportamento. Através da Figura 3.14, observamos que no ano 2019 ocorreu uma diminuição acentuada do número de palavras. Esta redução aconteceu porque neste ano não foram associados os anexos ao documento principal, ou seja, em 2019, os documentos têm apenas, maioritariamente, uma página. Nesta Figura, também é possível concluir que existe uma correlação direta entre as palavras e as frases.

Na Coleção existem 2.304 palavras diferentes, sendo que 1.977 (86%) pertencem ao dicionário de Língua Portuguesa e as restantes são números, marcas, abreviaturas ou palavras em inglês. Também foram corrigidas 128 palavras com erros ortográficos e removidas 70 palavras, fruto de erros do *OCR*.

As Entidades Mencionadas estão divididas em 3 níveis hierárquicos: Categoria, Tipo e Subtipo. A Categoria e o Tipo são os níveis obrigatório. A Figura 3.15 representa a distribuição das EMs na Coleção Dourada. Estão

representados os 11 tipos de Entidades Mencionadas mais frequentes, ou seja, mais de 15 ocorrências. Contudo, todas as diretivas das Entidades Mencionadas estão descritas mais à frente na Secção 4.2.

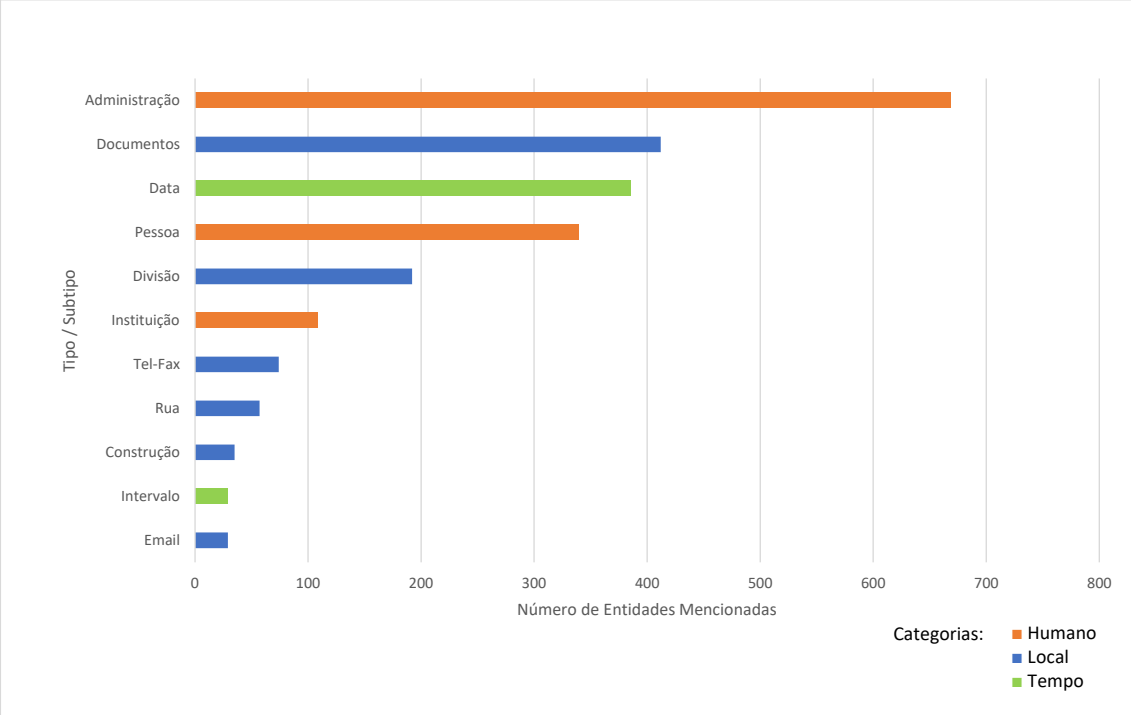

Figura 3.15: Distribuição das Entidades de Mencionadas na Coleção Dourada.

Os documentos normalizados, que são a maioria do *Corpus*, estão divididos em segmentos. Estes segmentos permitem compreender a distribuição dos dados na Figura 3.15. O Cabeçalho contém sempre a unidade da marinha (**Administração**), a cidade (**Divisão**) e data (**Data**). Por sua vez, nos segmentos Destinatários e Referências são identificados, respectivamente, as unidades (**Administração**) e os documentos (**Documentos**). No final de cada documento temos a morada (**Rua**) e o contacto (**Tel-Fax**).

Capítulo 4

Procedimentos e Implementação

Neste capítulo abordamos os procedimentos usados na anotação das entidades menciondas, assim como as diretivas utilizadas como guia para efectuar o reconhecimento, tendo em conta tanto a forma de identificação das entidades mencionadas como as categorias e tipos usadas na tarefa de classificação.

4.1 Procedimentos

Antes de anotar os documentos da Coleção Dourada, estes documentos tiveram uma fase de seleção e de limpeza. A fase de seleção consistiu em escolher, aleatoriamente, os documentos segundo a distribuição descrita na Secção 3.4. De seguida, foram submetidos a uma fase de limpeza, ou seja, as palavras incorretas foram corrigidas e os erros do *OCR* apagados. Ainda nesta fase, quando o final de uma frase não tinha um caracter de pontuação e o inicio da seguinte era uma letra em minúscula ou um número, estas duas frases foram concatenadas. Por fim, foi adicionado um ponto final no final de cada frase, quando este não existia, uma vez que a STRING só reconhece que é uma frase quando termina num caracter final, como o ponto final.

A tarefa de anotação da Coleção Dourada foi realizada segundo a seguinte metodologia. Em primeiro lugar, foram anotados 10 documentos, escolhidos aleatoriamente. De seguida, estes documentos foram submetidos a uma validação por um linguista com conhecimento tanto no domínio textual como nas diretivas de classificação usadas pela STRING, definidas na Secção 4.2. Por fim, foram anotados os restantes 200 documentos com base no *feedback* do linguista.

Esta metodologia foi escolhida de forma a minimizar o problema de falta de recursos, tanto de pessoas como de tempo. A metodologia que consideramos ser a mais correta, mas também a mais dispendiosa, é a seguinte: Primeiro, a Coleção Dourada seria anotada por 3 pessoas que compreendessem o domínio textual e as diretivas de classificação; De seguida, cada entidade mencionada seria discutida entre as 3 pessoas, prevalecendo a classificação da maioria. Desta forma, a anotação seria menos enviesada, pois, a decisão de cada um seria discutida em conjunto.

4.2 Diretivas

Nesta secção apresentam-se as diretivas para a identificação e classificação de entidades mencionadas baseadas no fórum de avaliação HAREM. Primeiro, realiza-se uma descrição do HAREM, de seguida, identifica-se os critérios gerais de identificação comuns a todos os tipos de entidades, e por fim, especifica-se os critérios de classificação para cada uma das seis categorias [13].

4.2.1 HAREM

O HAREM (*Avaliação de Reconhecimento de Entidades Mencionadas*) é um sistema orientado à língua portuguesa e baseia-se em regras manuscritas, tanto ao nível lógico (reconhecimento de padrões morfológicos) como global (contexto da frase), tendo como base uma gramática construtiva, que trata o reconhecimento de entidades mencionadas como uma tarefa integrante da anotação gramatical. As anotações das categorias candidatas são realizadas em três níveis e desambiguadas através de regras:

1. uso de entradas lexicais conhecidas e dicionários de termos (cerca de 17 000 entradas);
2. predição baseada em padrões morfológicos;
3. predição baseada no contexto para palavras que são desconhecidas.

4.2.2 Critérios de Identificação Geral

Os critérios de Identificação Geral usados pelo HAREM, também usados pela STRING, são os seguintes:

1. Uma EM deve conter pelo menos uma letra em maiúsculas, e/ou algarismos. No entanto, existe um conjunto de palavras relativas a certos domínios que também são excepções a esta regra, por exemplo, os meses do ano e o grau de parentesco (lista completa em [12]).
2. As frases totalmente escritas em maiúsculas (como acontece nos títulos dos documentos normalizados) devem ser analisados cuidadosamente, e só deverão conter etiquetas as EM claras.
3. Para evitar uma excessiva proliferação de EM com identificações alternativas, os sistemas e Coleção Dourada são construídos de forma a escolher a EM máxima, ou seja, aquela que contém, numa única interpretação possível, o maior número de palavras.

4.2.3 Diretivas de Classificação

De seguida é realizada uma descrição das diretivas de classificação usadas nos textos em português pela STRING [6][5]. As diretivas que não têm subtipo são descritas pelo seu tipo.

Categoria `VALOR`

Tem como objectivo capturar as várias entidades que aparecem nos textos como quantificadores numéricos.

- Tipo **`QUANTIDADE:`** este tipo pretende abranger as quantidades absolutas e relativas;
- Tipo **`BOLSA:`**
  - Subtipo **`MOEDA:`** este subtipo pretende capturar expressões que designem valores monetários.

Categoria `HUMANO`

- Tipo **`INDIVIDUAL:`**

  - Subtipo **PESSOA:** inclui os títulos e os graus de parentesco, que devem ser incluídos na EM que designa essa pessoa. Os diminutivos, alcunhas, iniciais, nomes mitológicos e entidades religiosas são também etiquetados nesta categoria;
  - Subtipo **CARGO:** Inclui um posto que pode ser desempenhado por diferentes pessoas ao longo do tempo. Os cargo que possuem na descrição uma organização, devem ter apenas uma etiqueta que abrange a organização. Por exemplo, diretor do Instituto Hidrográfico.
- Tipo **COLECTIVO:**
  - Subtipo **ADMINISTRAÇÃO:** inclui organizações relacionadas com a administração e governação de um território, tal como ministérios, secretarias de estado, municípios, câmaras e autarquias. Inclui também as organizações que têm a ver com a governação a nível internacional ou supra-nacional;
  - Subtipo **INSTITUIÇÃO:** todas as organizações que não possuem fins lucrativos (não sendo, portanto, empresas) nem um papel directo na governação são do tipo `INSTITUICAO`. Este tipo abrange instituições no sentido estrito, associações e outras organizações de espírito cooperativo, universidades, colectividades, escolas ou partidos políticos;
  - Subtipo **GRUPO:** esta diretiva abrange EM que se referem a um conjunto de pessoas como membros de uma organização ou conceito semelhante, tal como equipa ou seita.

Categoria LOCAL

- Tipo **CRIADO:**
  - Subtipo **PAÍS:** inclui países e uniões de países, como a União Europeia. Também inclui designações convencionais de certos países, tal como "País do Sol Nascente (Japão)"ou "Império do Meio (China)";
  - Subtipo **DIVISÃO:** Inclui agregados populacionais, tanto cidades, vilas e aldeias, como divisões administrativas, tal como estados no Brasil, municípios, distritos, províncias em Portugal ou regiões administrativas (Algarve);
  - Subtipo **CONSTRUÇÃO:** inclui todo o tipo de construções, desde edifícios ou áreas específicas de um edifício, até pontes, portos, arenas para eventos desportivos, etc;
  - Subtipo **RUA:** Inclui todos os tipos de estradas, ruas, avenidas, becos, praças, etc.
- Tipo **FÍSICO:**
  - Subtipo **AGUAMASSA:** inclui lagos, mares, oceanos, golfos, estreitos, canais, lagoas, etc.
- Tipo **VIRTUAL:**
  - Subtipo **SÍTIO:** inclui todas as localizações virtuais: WEB, WAP, FTP, etc;
  - Subtipo **DOCUMENTOS:** inclui regulamentos, leis, normas, decretos, diretrizes, protocolos, etc;
  - Subtipo **EMAIL:** inclui todos os endereços de mail;
  - Subtipo **TEL-FAX:** inclui todos os números de telefone, assim como os faxes e números de telemóvel, de acordo com os formatos válidos para vários países.

Categoria EVENTO

- Tipo **ORGANIZADO:**
  - Subtipo **POLITICO:** inclui eventos políticos, como eleições, congressos, marchas, etc.

  - Subtipo **OUTRO:** Engloba todos os eventos de diferentes áreas que não sejam política, desporto, ciência e arte.

Categoria DATUM

- Tipo **NIB:** inclui a entidade mencionada que representa os números de identificação bancária em Portugal;
- Tipo **MATRICULA:** abrange matrículas para veículos a motor, tal como os carros e as motas. Neste momento, as únicas matrículas que reconhece são as pertencentes a veículos portugueses.

Categoria TEMPO

As EMs de tipo TEMPO devem conter palavras que referem explicitamente a data ou a hora. Nota-se que, embora a idade de uma pessoa seja referida em anos é marcado como uma quantidade e não uma localização temporal.

- Tipo **TEMPO_CALEND:**
  - Subtipo **DATA:** inclui todas as referências a dias, mês e ano. Referências a mês e ano, ou só a ano, devem ser consideradas de tipo DATA se, no contexto, a referência indica uma localização temporal única;
  - Subtipo **INTERVALO:** engloba as EM que referem um intervalo de tempo contínuo e não repetido, com apenas um início e um fim;

Capítulo 5

Avaliação e Resultados

Neste capítulo descrevem-se os procedimentos utilizados na avaliação do sistema de reconhecimento e classificação de entidades mencionadas (Secção 5.1) e os resultados na Secção 5.2.1.

5.1 Avaliação

De forma a avaliar a tarefa de identificação e classificação de entidades mencionadas, recorreu-se a uma adaptação da metodologia utilizada pelo fórum de avaliação de RCEM da língua portuguesa, o HAREM [13]. Este fórum permite avaliar a correção dos resultados através do uso de uma coleção dourada, isto é, um documento de referência, em geral anotado manualmente, e que apresenta a saída ideal pretendida para a tarefa a avaliar.

A etiquetação do texto original, de acordo com as regras de etiquetagem da STRING, deve conter cada EM rotulada por uma etiqueta de abertura e de fecho, semelhante às etiquetas usadas em XML. Na etiqueta de abertura tem a categoria e o tipo atribuído, opcionalmente, também pode ter o subtipo.

O tipo ou subtipo é colocado entre aspas e tanto estes como a categoria devem estar em maiúsculas. Assim como, não devem existir espaços entre a entidade mencionada e as etiquetas que a rodeiam, e caracteres como aspas ou parênteses na parte etiquetada. Na Figura 5.1 apresenta-se um exemplo de uma etiquetação segundo o formato usado.

```
O <EM CATEG="HUMANO" TIPO="INDIVIDUAL" SUBTIPO="PESSOA">Sr. João</EM>
  foi à <EM CATEG="LOCAL" TIPO="CRIADO" SUBTIPO="PAÍS">Índia</EM>.
```

Figura 5.1: Exemplo de etiquetação de EMs de acordo com a STRING.

5.1.1 Medidas

Nesta subsecção são apresentadas as medidas usadas na tarefa de identificação e classificação de entidades mencionadas. No que diz respeito a tarefa de identificação, tem como objectivo medir eficiência do sistema e

delimitar as entidades de forma correta, em comparação com as entidades previamente anotadas existentes na coleção dourada.

O avaliador da tarefa de identificação atribui as seguintes classificações:

- **Correto:** Quando o elemento inicial e final da entidade mencionada são iguais na saída do sistema e na coleção dourada e o número total de elementos é igual entre si;
- **Parcialmente Correcto (por defeito):** Quando pelo menos um elemento de saída do sistema corresponde a um elemento de uma entidade mencionada na coleção dourada e o número total de elementos da entidade mencionada na saída do sistema é menor do que o número de elementos respectivos na coleção dourada;
- **Parcialmente Correto (por excesso):** Quando pelo menos um elemento de saída do sistema corresponde a um elemento de uma entidade mencionada na coleção dourada e o número total de elementos da entidade mencionada na saída do sistema é maior do que o número de elementos respectivos na coleção dourada;
- **Em Falta:** Quando o sistema falha a deteção correta de qualquer elemento de uma certa entidade mencionada presente na coleção dourada;
- **Espúrio:** Quando foi delimitada uma alegada entidade mencionada que não consta na coleção dourada, quer parcial ou totalmente

Enquanto às EMs classificadas como corretas é atribuído a pontuação de 1 e aos espúrios e entidades em falta a pontuação é de 0, as entidades mencionadas identificadas como parcialmente corretas são pontuadas segundo a equação (5.1):

$$\rho = 0,5\frac{n_c}{n_d} \tag{5.1}$$

Onde:

- n_c representa o número de elementos comuns entre a EM do sistema e a EM da CD, ou seja, a cardinalidade da intersecção dos elementos.
- n_d representa o número de elementos distintos entre a EM do sistema e a EM da CD, ou seja, a cardinalidade da reunião dos elementos.

A avaliação da classificação semântica tem como objetivo medir a capacidade do sistema em conseguir classificar uma entidade mencionada tendo em conta a hierarquia de categorias e tipos definidos pela STRING. A classificação semântica pode ser avaliada em quatro modalidades:

1. **classificação semântica por categorias:** apenas é considerada a categoria na etiqueta;
2. **classificação semântica combinada:** é avaliada tanto a correcção das categorias como dos tipos da entidade mencionada, através de uma pontuação que combina as duas;
3. **classificação semântica plana:** avalia-se os pares categoria-tipo, considerando apenas como certos os casos que tenham a categoria e o tipo pontuados como corretos.

No caso da classificação semântica combinada, a pontuação a atribuir é:

(i) 0, se a categoria não estiver correta;

(ii) 1, se a categoria estiver correta, mas o tipo estiver errado;

(iii) $1 + (1 - \frac{n_c}{n_t}) - \frac{n_e}{n_t}$ se a categoria estiver correta e pelo menos um dos tipos estiver correto, em que (n_c: o número de tipos corretos, n_e: o número de tipos espúrios, n_t: o número de tipos possível nessa categoria);

A precisão mede a qualidade de resposta do sistema que mede a proporção de respostas corretas em relação a todas as respostas dadas pelo sistema. Na tarefa de identificação, a precisão mede a relação entre as entidades corretas e parcialmente corretas de todas as entidades identificadas pelo sistema, e é calculada de acordo com a fórmula (5.2) em que x é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada.

$$Precisão_{\,Identificação} = \frac{Num \quad EMs \quad Corretas \quad + \quad x}{Num \quad EMs \quad Identificadas} \tag{5.2}$$

Em relação à classificação semântica, há que ter em conta as quatro modalidades descritas anteriormente: classificação por categorias, classificação por tipo, classificação semântica combinada e classificação semântica plana.

No que diz respeito à classificação por categorias, o cálculo da precisão está definido na fórmula (5.3) em que y é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com a categoria correta.

$$Precisão_{\,Classif.\ por\ Catergorias} = \frac{(Num \quad EMs \quad com \quad Identificação \quad e \quad Categoria \quad Correta \quad + \quad y)}{Num \quad EMs \quad Classificadas} \tag{5.3}$$

Para a classificação semântica combinada, a precisão mede o grau de sucesso de acordo com a classificação máxima (calculada assumindo que todas as categorias e tipos propostos pelo sistema estão corretos) e é dada pela fórmula (5.4).

$$Precisão_{\,Classif.\ Combinada} = \frac{Valor \quad Classif.Semântica \quad do \quad Sistema}{Valor \quad Máximo \quad Classif.Semântica \quad p/Saída \quad do \quad Sistema} \tag{5.4}$$

No caso da classificação plana, a precisão é dada pela fórmula (5.5) em que z é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com categoria e tipo corretos.

$$Precisão_{\,Classif.\ Plana} = \frac{(Num \quad EMs \quad com \quad Identificação, \quad Categoria \quad e \quad Tipo \quad Correto \quad + \quad z)}{Num \quad EMs \quad Classificadas} \tag{5.5}$$

A abrangência (ou cobertura) mede a percentagem de respostas correctas que o sistema conseguiu identificar. Na tarefa de identificação, a abrangência mede a quantidade de entidades mencionadas da coleção dourada que foram identificadas e é dada pela fórmula (5.6) em que x é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada.

$$Abrangência_{\,Identificação} = \frac{Num \quad EMs \quad Corretas \quad + \quad x}{Num \quad EMs \quad na \quad Coleção \quad Dourada} \tag{5.6}$$

De modo similar ao cálculo da precisão, a abrangência para a classificação semântica é definida de maneira diferente para cada uma das modalidades de avaliação. O cálculo da abrangência no caso da avaliação por categorias é dado pela fórmula (5.7) em que y é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com a categoria correta.

$$Abrangência_{\ Classif.\ por\ Catergorias} = \frac{(Num\quad EMs\quad com\quad Identificação\quad e\quad Categoria\quad Correta + y)}{Num\quad EMs\quad na\quad Coleção\quad Dourada} \tag{5.7}$$

Na avaliação da classificação semântica combinada, a abrangência mede o nível de cobertura de acordo com a classificação máxima (se tanto as categorias como os tipos enviados estiverem corretos) e é dada pela fórmula (5.8).

$$Abrangência_{\ Classif.\ Combinada} = \frac{Valor\quad Classif.Semântica\quad do\quad Sistema}{Valor\quad Máximo\quad Classif.Semântica\quad na\quad Coleção\quad Dourada} \tag{5.8}$$

Por fim, relativamente à classificação plana, a abrangência é dada pela fórmula (5.9) em que z é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com categoria e tipo corretos.

$$Abrangência_{\ Classif.\ Plana} = \frac{(Num\quad EMs\quad com\quad Identificação,\quad Categoria\quad e\quad Tipo\quad Correto + z)}{Num\quad EMs\quad na\quad Coleção\quad Dourada} \tag{5.9}$$

A medida-F combina as medidas de precisão e de abrangência para cada tarefa, de acordo com a fórmula (5.10).

$$Medida - F = \frac{(2 * Precisão * Abrangência)}{Precisão + Abrangência} \tag{5.10}$$

A sobregeração mede o excesso de resultados espúrios que um sistema produz, ou seja, quantas vezes produz resultados errados. Relativamente à tarefa de identificação, a sobregeração mede quantas entidades mencionadas identificadas pelo sistema não existem na coleção dourada e é calculada através da fórmula (5.11).

$$Sobregeração_{\ Identificação} = \frac{Num\quad de\quad EMs\quad espúrias}{Num\quad de\quad EMs\quad identificadas} \tag{5.11}$$

A sobregeração na classificação semântica mede o número de entidades mencionadas com uma classificação semântica espúria, em comparação com a coleção dourada. No caso de avaliação por categorias, a sobregeração é dada pela fórmula (5.12)

$$Sobregeração_{\ Classif.\ por\ Catergorias} = \frac{Num\quad de\quad EMs\quad espúrias\quad na\quad Categoria}{Num\quad de\quad EMs\quad classificadas\quad na\quad Categoria} \tag{5.12}$$

No caso da classificação plana, a sobregeração é calculada segundo as fórmula (5.13)

$$Sobregeração_{\ Classif.\ Plana} = \frac{Num\quad de\quad EMs\quad espúrias\quad na\quad Categoria\quad ou\quad Tipo}{Num\quad de\quad EMs\quad classificadas\quad na\quad Categoria\quad e\quad Tipo} \tag{5.13}$$

A subgeração é uma medida de quanto faltou ao sistema analisar, dada a solução conhecida, e.g. a coleção dourada. Para a tarefa de identificação, mede a quantidade de entidades mencionadas que existem na coleção dourada que não foram identificadas pelo sistema e é calculada através da fórmula (5.14).

$$Subgeração_{Identificação} = \frac{Num\ de\ EMs\ em\ Falta}{Num\ de\ EMs\ na\ Coleção\ Dourada} \tag{5.14}$$

A subgeração na classificação semântica mede as classificações semânticas em falta. No caso da avaliação por categorias, a subgeração é calculada de acordo com a fórmula (5.15).

$$Subgeração_{Classif.\ por\ Catergorias} = \frac{Num\ de\ EMs\ em\ Falta\ na\ Categoria}{Num\ de\ EMs\ classificadas\ na\ Categoria} \tag{5.15}$$

Por último, no que diz respeito à avaliação plana, a subgeração é calculada de acordo com a fórmula (5.16).

$$Subgeração_{Classif.\ Plana} = \frac{Num\ de\ EMs\ em\ Falta\ no\ Tipo}{Num\ de\ EMs\ Classificadas\ na\ Categoria\ na\ Coleção\ Dourada} \tag{5.16}$$

5.2 Resultados

Nesta Secção do documento apresentamos os resultados obtidos pela cadeia de processamento de Língua Natural durante a avaliação da tarefa de Reconhecimento e Classificação de Entidades Mencionadas. Na Subsecção 5.2.1 são descritos os resultados gerais. Devido ao resultado geral pouco convincente, consideramos relevante segmentar o resultado nas diferentes categorias, de forma a perceber a razão desse resultado. Esta análise está presente na Subsecção 5.2.2.

5.2.1 Resultado Geral

A avaliação foi realizada através da comparação entre os documentos anotados da Coleção Dourada e o *output* da tarefa de RCEM da STRING. O resultado da identificação está apresentado na tabela 5.1.

	Precisão	Abrangência	Medida-F	Sobregeração	Subgeração
Identificação	41,39%	55,41%	47,39%	25,30%	26,80%

Tabela 5.1: Resultados da Identificação no Geral.

Como referido anteriormente, a precisão mede a relevância do resultado. A precisão da tarefa de Identificação das Entidades Mencionadas foi de 41,39%, pelo que concluímos que, a relevância do resultado é baixa.

A abrangência mede a quantidade de resultados relevantes que foram devolvido. Este valor foi mais elevado do que a precisão, contudo o valor mantêm-se baixo, 55,41%.

Para medir o balanço entre a precisão e abrangência usamos a Medida-F. Esta medida é calculada apenas em função da precisão e da abrangência através da Equação 5.10. Assim, a medida-F descreve a fiabilidade do resultado, que neste caso é baixo.

A medida usada para avaliar o número de entidades mencionadas espúrias. ou seja, identificadas pela STRING mas ausentes da Coleção Dourada, é a sobregeração. O valor desta medida é discutido no parágrafo sobre a categoria *Valor*.

Por outro lado, a subgeração mede o número de entidades mencionadas que não foram identificadas pela cadeia de processamento de Língua Natural. Este valor é descrito ao longo das categorias da Subsecção 5.2.2.

Classificação	Precisão	Abrangência	Medida-F	Sobregeração	Subgeração
Por Categorias	86,42%	55,41%	67,53%	52,81%	55,94%
Combinada	83,72%	53,68%	65,42%	-	-
Plana	67,96%	43,58%	53,10%	33,87%	26,80%

Tabela 5.2: Resultados da Classificação no Geral.

Os resultados gerais da Classificação estão apresentados na Tabela 5.2. A Precisão para a Classificação semântica por Categoria e Combinada é elevada, contudo, quando temos em conta o par Categoria-Tipo a precisão baixa.

Tal como na tarefa de Identificação, a Abrangência dos resultados é baixa. A Sobregeração e a Subgeração têm valores demasiado elevados. Isto acontece devido a problemas da tarefa de RCEM da STRING. Estes problemas são descritos na próxima Subsecção 5.2.2.

5.2.2 Resultado Discriminado

Nesta Subsecção, analisamos os resultados discriminados para as três categorias mais representativas, descritas no Capítulo 4. Apesar de a categoria *Valor* não ser uma categoria representativa, considerámos relevante para descrever os espúrios usados na sobregeração. De seguida, são apresentados os resultados obtidos, assim como os problemas que impediram a obtenção de melhores resultados.

Humano

Os resultados de identificação da categoria *Humano* são apresentados na Tabela 5.3.

	Precisão	Abrangência	Medida-F	Sobregeração	Subgeração
Identificação	50,30%	57,83%	53,80%	13,02%	27,87%

Tabela 5.3: Resultados da Identificação na Categoria *Humano*.

O principal problema nesta categoria foi a falta de palavras do domínio da Marinha no vocabulário da STRING. Trata-se de palavras como, *Superintendência*, *Comando Naval* e *Escola de Tecnologias Navais*. Estas palavras são frequentes no *Corpus*, contudo nunca foram identificadas como Entidades Mencionadas pela STRING.

Outro problema foi as abreviaturas: nenhuma abreviatura da unidade ou da patente foi reconhecida como EM, por não constar na lista de abreviaturas da STRING.

A precisão da classificação por categorias e combinada são elevadas. Isto acontece porque a STRING classifica a maioria das palavras em maiúsculas como pessoas, ou seja, com a Categoria *Humano* e Tipo *Individual*. Na

Classificação	Precisão	Abrangência	Medida-F	Sobregeração	Subgeração
Por Categorias	85,10%	57,83%	68,86%	22,02%	47,15%
Combinada	99,20%	67,42%	80,28%	-	-
Plana	47,02%	31,95%	38,05%	14,97%	27,87%

Tabela 5.4: Resultados da Classificação na Categoria *Humano*.

Marinha, as palavras em maiúsculas são usadas para as Instituições e órgãos da Administração, logo, a STRING acerta na Categoria destas Entidades Mencionadas.

Por outro lado, a abrangência continua com maus resultados. A subgeração também tem um valor demasiado elevado. As principais razões para este resultado são os problemas identificados na tarefa de identificação das EMs para a categoria *Humano*, que vimos acima.

De seguida, iremos apresentar alguns exemplos das EMs reconhecidas, parcialmente reconhecidas e não reconhecidas pela STRING. Para esta categoria, a cadeia PLN reconheceu corretamente as seguintes EMs: "*Escola Naval*", "*Instituto Hidrográfico*"e "*Gabinete do Chefe do Estado-Maior da Armada*". Adicionalmente, reconheceu as seguintes EMs como parcialmente corretas: "*Direção Geral da Autoridade Marítima*", "*Direção de Análise e Gestão da Informação*"e "*Inspecção-Geral da Marinha*". Por fim, não reconheceu as seguintes EMs: "*Comando Naval*", "*Escola de Tecnologias Navais*"e "*Superintendente dos Serviços de Tecnologias da Informação*".

Local

Para a categoria *Local*, os resultados da identificação das EMs são apresentados na Tabela 5.5.

	Precisão	Abrangência	Medida-F	Sobregeração	Subgeração
Identificação	40,55%	43,65%	42,05%	7,08%	43,99%

Tabela 5.5: Resultados da Identificação na Categoria *Local*.

A Medida-F tem um valor baixo, pois está diretamente dependente da precisão e da abrangência. Estas medidas têm valores baixos porque a STRING não reconhece certas palavras como documentos, tais como Relatório, Despacho ou Diário. Estas palavras são Entidades Mencionadas com a Categoria *Local* e Tipo *Virtual*, normalmente, presentes no Segmento *Referência* do documento normalizado.

Classificação	Precisão	Abrangência	Medida-F	Sobregeração	Subgeração
Por Categorias	82,89%	43,65%	57,18%	14,47%	89,91%
Combinada	93,72%	48,98%	64,17%	-	-
Plana	82,67%	43,53%	57,03%	7,62%	43,99%

Tabela 5.6: Resultados da Classificação na Categoria *Local*.

Semelhante à categoria *Humano*, esta categoria também tem bons resultados na precisão. Isto acontece porque a STRING tem facilidade em reconhecer Entidades Mencionadas do subtipo *Rua*, pois estas EMs, normalmente, contêm a própria palavra-chave que permite identificar esta categoria (*rua*, *avenida*, *praça*, etc).

Porém, a percentagem de Entidades Mencionadas que não foram reconhecidas para esta categoria é elevado, 89,91%, pois, como referido anteriormente, a STRING tem problemas em identificar os documentos, assim como os locais específicos deste ramo das Forças Armadas. Trata-se do problema do vocabulário limitado, que vimos no parágrafo acerca da categoria *Humano*.

Para a categoria *Local*, a STRING reconheceu corretamente as seguintes palavras: "*Lisboa*", "*dsf@marinha.pt*"e "*Atlântico Sul*". Por outro lado, reconheceu parcialmente as palavras: "*V/Nota n.º 17/DAF*", "*Proposta n.º 32*"e "*Decreto-Lei nº172/94*". Por fim, não reconheceu as seguintes EMs: "*Relatório com a entrada nº 1376*", "*ITSUF 2*"e "*Decreto Regulamentar n.º 24/94*".

Tempo

Para a categoria *Tempo*, os resultados da identificação estão representados na Tabela 5.7.

	Precisão	Abrangência	Medida-F	Sobregeração	Subgeração
Identificação	62,12%	71,61%	66,53%	13,25%	20,07%

Tabela 5.7: Resultados da Identificação na Categoria *Tempo*.

Com esta tabela, concluímos que a STRING tem dificuldades em identificar EMs referentes ao tempo, como as datas. Isto acontece porque existem datas que são representadas como abreviaturas e a STRING não está preparada para reconhecê-las.

Classificação	Precisão	Abrangência	Medida-F	Sobregeração	Subgeração
Por Categorias	93,18%	71,61%	80,99%	2,00%	30,11%
Combinada	84,27%	64,77%	73,24%	-	-
Plana	62,12%	71,61%	80,98%	15,28%	25,12%

Tabela 5.8: Resultados da Classificação na Categoria *Tempo*.

Apesar de a identificação das EMs desta categoria não ser muito positiva, a tarefa de classificação tem bons resultados, pois, na maioria das vezes que uma Entidade Mencionada da Categoria *Tempo* é identificada, a sua classificação é realizada de forma correta. Isto acontece porque, em Portugal, os elementos que formam as datas seguem, geralmente, uma ordem específica, primeiro o dia, seguido do mês, e, por fim, o ano.

Para a categoria *Tempo*, a cadeia de PLN reconheceu corretamente as seguintes palavras: "*de 25 de julho de 2014*", "*1º quadrimestre de 2014*"e "*de 07 de Abril*". Por outro lado, reconheceu parcialmente as palavras: "*de 27 a 30 de outubro de 2014*", "*"inicio do 2.º semestre de 2016*"e "*11 dias de férias*". Por fim, não reconheceu as seguintes EMs: "*1.º QUADRIMESTRE DE 2014*", "*46MAI15*"e "*de 03JUN1992 a 18ABR1994*".

Valor

A categoria *Valor* tem pouca representatividade (0,54%) na Coleção Dourada, contudo tem problemas na tarefa de RCEM da STRING, pelo que influenciam o resultado geral. Foram identificados os seguintes problemas no reconhecimento de:

- código postal;

- abreviaturas das datas, por exemplo, a data *27MAI14* que representa a data 27 de Maio de 2014;
- número de identificação dos militares;
- números de identificação dos documentos
- eventos, por exemplo, o evento *II Encontro Anual da I&D em Ciências Militares*.

Todos os problemas identificados são do tipo *Quantidade*. Como são erros da tarefa de RCEM, são identificados como espúrios, contribuindo para a sobregeração do resultado geral. Esta categoria é relevante porque contém 64% do número total de espúrios.

Capítulo 6

Conclusão e Trabalho Futuro

Durante o processamento do *Corpus*, reparámos que a principal dificuldade surge pela leitura inábil do *OCR*. Um Trabalho Futuro poderá ser o desenvolvimento de uma ferramenta OCR que consiga identificar a informação desnecessária, como imagens e selos, e as diferentes secções do documento.

Na Marinha Portuguesa, o departamento de Gestão de Recursos está dividido em 4 superintendências, enquanto o *Corpus* apenas contempla a superintendência da Tecnologia de Informação. Logo, um Trabalho Futuro poderá estender o trabalho desenvolvido para os documentos das outras superintêndencias.

Outro Trabalho Futuro poderá ser variar o domínio específico do *Corpus*. Neste caso, o domínio é a correspondência da Marinha, contudo existem organizações semelhantes, tais como um ramo das Forças Armadas (e.g. Força Aérea) ou um departamento do Governo de Portugal (e.g. Ministério da Defesa), que poderão ser de interesse para aplicar a estratégia utilizada por este projecto.

Tendo como ponto de partida o *Corpus* processado por este projeto, seria interessante desenvolver duas soluções de classificação, uma usando o número do processo e outra a tabela de distribuição. O número do processo tem informação acerca do tipo de documento, por outro lado, a tabela distribuição referencia os destinatários do documento. Contudo, os dois classificadores complementam-se para completar a tarefa de distribuição automática de documentos entre departamentos. Esta tarefa tem como objetivo optimizar o fluxo de informação e reduzir o erro e esforço humano.

O número de processo é um classificador do documento, logo o projeto que o usará terá o objetivo de classificar o documento através do seu conteúdo. Para este projeto, o *Corpus* contém 4.737 documentos. Como não são muitos documentos devem ser ponderadas três abordagens. A primeira é o uso de métodos tradicionais de classificação, como o KNN e SVM. A segunda é a incorporação pré-treinada de palavras. Estas palavras seriam inseridas em modelos RNN e CNN para capturar o significado dos documentos, por fim, a terceira abordagem seria usar redes neuronais para treinar o modelo, esta abordagem usaria o modelo hierárquico e a incorporação de pré-treinada palavras. Por outro lado, o projeto que usará a tabela de distribuição terá como objetivo estudar a eficiência dos métodos de classificação de multi-etiquetas na tarefa de distribuição de documentos. Este projeto não tem um classificador bem definido, como o projecto anterior, logo o uso de métodos simples poderá ser a solução para combater o caráter disperso do classificador.

Um dos objetivos deste projecto era a adaptação da cadeia de PLN para um domínio textual específico. Este objectivo não foi atingido porque houve a necessidade de melhorar o *Corpus* para a tarefa de aprendizagem automática. Pois, os dois projetos que procedem este projeto foram desenvolvidos ao mesmo tempo que este. Apesar de o Reconhecimento e Classificação de Entidades Mencionadas não ter sido melhorado, o *Corpus* desenvolvido permitiu avaliar a STRING para este domínio específico. Assim, para um futuro trabalho de melhoria da STRING, já existe um *Corpus* preparado e a identificação dos principais problemas da cadeia de PLN.

Algumas melhorias a este projecto seria adicionar ao vocabulário da STRING as palavras desconhecidas, assim como as abreviaturas. De seguida, partindo da avaliação realizada, identificar as EMs que a STRING não conseguiu reconhecer ou reconheceu parcialmente corretas. Por fim, adicionar as regras necessárias para que a STRING passe a reconhecer e classificar corretamente as Entidades Mencionadas. Depois de aplicadas estas melhorias, seria interessante tornar a avaliar a tarefa de RCEM da STRING e comparar com este projecto. Desta forma, concluíamos acerca das melhorias realizadas.

O processamento do *Corpus* tornou-se uma tarefa árdua e desafiante, pois os documentos tinham muita informação desnecessária que foi necessário remover; os documentos eram muito diversos, tanto quanto ao tipo documental como à língua usada; o facto de este *Corpus* ser usado para duas classificações diferentes tornou necessário um tratamento diferenciado; por fim, e muito significativo, o facto de o *OCR* não ser especializado para documentos em português dificultou o correto reconhecimento de palavras, principalmente palavras com acentos.

Este projeto é a primeira parte de duas para realizar a tarefa de aprendizagem automática nos documentos da Marinha, por isso os documentos foram sujeitos a um tratamento rigoroso para garantir que os dados do *Corpus* eram válidos e autênticos. Deste modo, os dois projetos seguintes terão a matéria-prima necessária para o sucesso na tarefa de aprendizagem automática.

Referências

[1] Decreto-Lei n.º 148/2015. *D.R. I Série*, 1 (2015-07-31),(148):19–21.

[2] Decreto-Lei nº 250/2014. *D.R. I Série*, 1 (2014/09/29),(250):6397–6406.

[3] Babych, B. and Hartley, T. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*, pages 1–8.

[4] Baptista, J., Mamede, N., and Gomes, F. (2010). Auxiliary verbs and verbal chains in European Portuguese. In *Computational Processing of the Portuguese Language*, pages 110–119, Berlin, Heidelberg. Springer Berlin Heidelberg.

[5] Baptista, J., Mamede, N., Hagège, C., and Maurício, A. (2012). Guidelines for identication, classication and normalization. Technical report, INESC-ID, Lisboa, Portugal.

[6] Baptista, J., Mamede, N., Oliveira, D., and Santos, D. (2011). Classication directives for named entities in portuguese texts. Technical report, INESC-ID, Lisboa, Portugal.

[7] Brun, C. and Hagège, C. (2004). Intertwining deep syntactic processing and named entity detection. In *Advances in Natural Language Processing*, volume 3230, pages 195–206, Alicante, Spain. Springer.

[8] Cabrita, V. (2013). Identificar, Ordenar e Relacionar Eventos. Master's thesis, Instituto Superior Técnico.

[9] Diniz, C. (2010). RuDriCo2 - Um Conversor Baseado em Regras de Transformação Declarativas. Master's thesis, Instituto Superior Técnico, Lisboa.

[10] Diniz, C., Mamede, N., and Pereira, J. (2010). RuDriCo2 - A Faster Disambiguator and Segmentation Modifier. In *INFORUM II*, pages 573–584.

[11] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

[12] e Diana Santos, N. C. (2006). Directivas e categorias para identificação e classificação semântica na coleccção dourada do harem. In *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, pages 211–239.

[13] Equipa da Linguateca (2019). HAREM: Reconhecimento de entidades mencionadas em português. https://www.linguateca.pt/HAREM/. Último acesso 2019-10-24.

[14] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

[15] Mamede, N., Baptista, J., Diniz, C., and Cabarrão, V. (2012). STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In *PROPOR 2012*, volume Demo Session.

[16] Mandl, T. and Womser-Hacker, C. (2005). The effect of named entities on effectiveness in cross-language information retrieval evaluation. *Proceedings of the ACM Symposium on Applied Computing*, 2:1059–1064.

[17] Marinha Portuguesa (2019). Estrutura da Marinha. https://www.marinha.pt/pt/a-marinha/Paginas/estrutura.aspx. Último acesso 2019-05-08.

[18] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482 – 489.

[19] Maurício, A. (2011). Identificação, classificação e normalização de expressões temporais. Master's thesis, Instituto Superior Técnico, Lisboa.

[20] Medeiros, J. C. (1995). Processamento Morfológico e Correção Ortográfica do Português. Master's thesis, Instituto Superior Técnico, Lisboa.

[21] Media4x (2020). Most Common English Words. https://www.rypeapp.com/most-common-english-words/. Último acesso 2020-05-31.

[22] NetOwl (2017). When 80% of the world's data is unstructured, entity extraction is a must. https://www.netowl.com/2017/08/11/80-worlds-data-unstructured-entity-extraction-must. Último acesso 2019/03/26.

[23] Nobata, C., Sekine, S., Isahara, H., and Grishman, R. (2002). Summarization system integrated with named entity tagging and ie pattern discovery. pages 1742–1745.

[24] Nobre, N. (2011). Resolução de expressões anafóricas. Master's thesis, Instituto Superior Técnico, Lisboa.

[25] Oliveira, D. (2010). Extraction and Classification of Named Entities. Master's thesis, Instituto Superior Técnico, Lisboa.

[26] Pizzato, L. A., Molla, D., and Paris, C. (2006). Pseudo relevance feedback using named entities for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 83–90, Sydney, Australia.

[27] Reinsel, D., Gantz, J., and Rydning, J. (2008). The Digitization of the World - From Edge to Core. Technical report, Seagate, United States.

[28] Ribeiro, R. (2003). Anotação morfossintática desambiguada do português. Master's thesis, Instituto Superior Técnico, Lisboa.

[29] Ribeiro, R., Oliveira, L., and Trancoso, I. (2003). Using morphossyntactic information in TTS Systems: Comparing strategies for European Portuguese. In *PROPOR 2003 - Computational Processing of the Portuguese Language: 6th International Workshop*, pages 143–150.

[30] Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2019). HAREM: An Advanced NER Evaluation Contest for Portuguese. pages 1986–1991.

[31] Sekine, S. and Isahara, H. (2000). IREX: IR and IE Evaluation project in Japanese. In *Proceedings of International Conference on Language Resources Evaluation (LREC 2000)*.

[32] Shilakes, C. and Tylman, J. (1998). Enterprise Information Portals. Technical report, Merrill Lynch, United States.

[33] Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

[34] Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

[35] Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Apêndice A

Categorias Morfossintáticas

Nome	Categoria	Exemplos
Adj	Adjectivo	espirituoso
Adv	Advérbio	ontem, amanhã
Art	Artigo	o, a
Conj	Conjunção	e, ou
Foreign	Palavra estrangeira	court
Interj	Interjeição	ui
Noun	Nome comum ou próprio	boca, Nuno
Num	Numérico	quinze, 1904, primeiro
Pastpart	Participio passado	lavado, amado
Prep	Preposição	em, para, com
Pron	Pronome	ele, meu, este, algo
Punct	Pontuação	;, :, .
Rel	Pronome relativo	qual, que
Symbol	Símbolo especial	$, %, #
Verb	Verbo	comi, andaram

Tabela A.1: XIP: lista de categorias morfossintáticas.