INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Identificação, Classificação e Normalização de Expressões Temporais

Andreia Sofia Baptista Maurício

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática e de Computadores

Presidente:	Professor Doutor João António Madeiras Pereira
Orientador:	Professor Doutor Nuno João Neves Mamede
Co-Orientador:	Professor Doutor Jorge Manuel Evangelista Baptista
Vogal:	Professor Doutor Bruno Emanuel da Graça Martins

Novembro de 2011

Agradecimentos

Gostaria de agradecer ao meu orientador, o Professor Nuno Mamede, e ao meu co-orientador Professor Jorge Baptista, a sua constante disponibilidade na resolução de todos os problemas que surgiram no desenvolvimento da dissertação, só com o seu apoio, saber e experiência me foi possível escrever estas linhas finais de agradecimento que dificilmente reflectem a gratidão que sempre recordarei.

Um obrigado especial aos meus pais, avós, à minha irmã e ao Rui que sempre me apoiaram nesta difícil tarefa. Em especial aos meus pais, que nos momentos menos bons deste percurso me deram todo o apoio e paciência e à minha irmã pelos exemplos de tenacidade e resistência que nos dá todos os dias, mostrando que o "impossível"não existe.

Aos meus colegas Daniel Santos, Diogo Oliveira, Ricardo Portela e ao Tiago Simão o meu muito obrigado, por todos os momentos que partilhámos juntos. Momentos de trabalho, de ajuda, de riso, de tristeza, ..., muitos dos momentos que convosco vivi, serão recordados pela minha vida fora.

Termino com um agradecimento muito grande a todos os meus familiares e amigos que de uma forma ou de outra, por vezes sem o saberem, me ajudaram nesta difícil tarefa.

Lisboa, Novembro de 2011

Andreia Sofia Baptista Maurício

Para a minha irmã Daniela.

Resumo

Enquadrado no âmbito mais alargado do REM, esta dissertação tem como objectivo identificar classificar e normalizar expressões temporais que ocorrem num texto escrito.

As directivas adoptadas para a classificação e normalização das expressões temporais surgem no seguimento do trabalho desenvolvido para o Segundo HAREM. As directivas foram corrigidas e adaptadas, tendo em conta algumas das dificuldades sentidas e as limitações apontadas a esta avaliação conjunta.

No âmbito desta dissertação foi desenvolvido um módulo de processamento de expressões temporais que está integrado na cadeia de processamento do L2F, no INESC-ID em Lisboa, que se apresenta detalhadamente neste documento.

Adicionalmente, neste documento também é realizada uma análise e comparação das metodologias utilizadas por outros sistemas que identificam expressões temporais.

Abstract

As part of the REM task, this dissertation aims to identify, classify and normalize temporal expressions contained in a written text.

The TIMEX guidelines adopted for the classification and normalizations of Time Expressions resulted from the participation in the Second HAREM Joint Evaluation Campaign. These guidelines were extended and adapted to identify more complex types of TIMEX.

As part of this dissertation, a processing module for time expressions was developed, that is integrated in the processing of the L2F at INESC-ID in Lisbon. In this document the processing module will be presented in detail.

Additionally, this document also contains an analysis and comparison of the methodologies used by other systems that also identify temporal expressions.

Palavras Chave
Keywords

Palavras Chave

Expressões Temporais

Reconhecimento de Entidades Mencionadas

Sistema de Processamento de Linguagem Natural

Classificação

Identificação

Normalização

Directivas

Keywords

Time Expressions

Named Entity Recognition

Natural Language Processing System

Identification

Classification

Normalization

Guidelines

Índice

Lista de Figuras

Lista de Tabelas

1 Introdução

O Reconhecimento de Entidades Mencionadas (REM) é uma sub-tarefa da área de Extração de Informação e tem como objectivo localizar e classificar diversas expressões linguísticas que designam diferentes tipos de entidades no universo extra-linguístico (pessoas, cargos, locais, etc.) num texto escrito. A sua importância prende-se com o facto do REM ser um componente necessário para sistemas de extração de informação, de resposta automática a perguntas, de tradução automática ou de sumarização de texto.

Em 2002, a Linguateca[1] reuniu esforços para avaliar as sub-áreas do processamento computacional do português. Até aí, os trabalhos publicados sobre português em geral apenas apresentavam a sua auto-avaliação, aspecto que impedia uma comparação directa entre as diversas abordagens e metodologias. Por este motivo, a Linguateca organizou uma Avaliação de Reconhecedores de Entidades Mencionadas: o HAREM[2], cujo objectivo era a comparação do desempenho de diversos sistemas usando um conjunto de recursos comuns e métricas consensuais. Ao comparar as diversas abordagens, o HAREM pretendia estimular o desenvolvimento dos sistemas e contribuir para o aperfeiçoamento do seu desempenho.

A primeira edição do HAREM realizou-se em 2005, e aí participaram dez sistemas. A segunda edição decorreu em 2008 e teve a participação do mesmo número de sistemas, entre estes encontrava-se o sistema "Reconhecimento de Entidades Mencionadas com o XIP" que resultou de uma colaboração entre o L^2F do INESC-ID Lisboa[3] e a Xerox. Este sistema apoia-se numa cadeia de processamento em que certos módulos foram desenvolvidos pelo L^2F e o parser (analisador sintáctico) XIP (Xerox Incremental Parsing) integrando no final da cadeia. É este último módulo que procede à identificação e classificação das entidades mencionadas.

Esta participação apoia-se igualmente em trabalhos anteriores desenvolvidos no âmbito do

[1] http://www.linguateca.pt.
[2] http://www.linguateca.pt/HAREM .
[3] Laboratório de Sistemas de Língua Falada do Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento em Lisboa.

L^2F, nomeadamente no reconhecimento de expressões temporais (Loureiro 2007) e de outras categorias de entidades (Romão 2007). Mais recentemente outros trabalhos vieram juntar-se aos referidos anteriormente, nomeadamente no que respeita às entidades do tipo valor, pessoa (Oliveira 2010), à extração de relações entre entidades (Santos 2010) e à resolução da referência de expressões anafóricas (Nobre 2011).

1.1 Motivação

Enquadrado no âmbito mais alargado do REM, este trabalho tem como objectivo identificar, classificar e normalizar expressões temporais (ET) que ocorrem num texto escrito. Assim pretende-se localizar e classificar as palavras de um texto que se referem a datas, durações, intervalos temporais e frequências.

A principal motivação para a identificação e normalização das ET, consiste na recolha de informação para uma ordenação cronológica e eventos, o que permitirá que os diversos eventos mencionados num texto possam ser ordenados numa sequência cronológica. Com a normalização das ET ambiciona-se a uniformização das mesmas para um formato-padrão, que poderá ser, posteriormente, utilizado por diferentes sistemas.

As directivas seguidas para a classificação e normalização de ET encontram-se descritas na proposta *"Time Expression in Portuguese: Guidelines for Identification, Classification and Normalization"* (Baptista, Mamede, Hagège, & Maurício 2009). Estas novas directivas surgem no seguimento do trabalho desenvolvido para o Segundo HAREM e resultam da correcção e adaptação a diversas dificuldades sentidas, bem como de algumas limitações apontadas a essa avaliação conjunta. Visa-se, desta forma, classificar e normalizar um número mais abrangente de ET.

1.2 Estrutura do Documento

Este documento está dividido em 6 capítulos. No primeiro capítulo é feita a introdução ao tema e é apresentada a motivação desta dissertação.

No segundo capítulo são abordados os trabalhos relacionados com esta dissertação. É feita uma descrição dos sistemas que participaram no Segundo HAREM e das diferentes abordagens

seguidas por estes sistemas. Adicionalmente, são também descritas as directivas seguidas no segundo HAREM para a categoria TEMPO, assim como o diferente envolvimento nos sistemas com o TEMPO e os seus resultados na avaliação. Ainda neste capítulo, é abordada a cadeia de processamento do L2F e os diferentes módulos que a constituem.

As novas directivas são apresentadas no Capítulo 3, onde são focadas as principais diferenças relativamente às directivas no Segundo HAREM.

O processo de implementação desta dissertação é descrito no Capítulo 4. Na primeira parte deste capítulo é descrito o processo de implementação na identificação e classificação de expressões temporais, que se realiza na cadeia de processamento do L2F. Na segunda parte, é descrito o processo de implementação do módulo de pós-processamento, responsável pela normalização das expressões temporais.

No Capítulo 5, descreve-se a avaliação do módulo de processamento de ET desenvolvido nesta dissertação, a metodologia utilizada e os resultados obtidos.

No último capítulo, o Capítulo 6, são feitas sugestões para o trabalho futuro a nível da classificação e normalização de expressões temporais. Para finalizar, é feita uma reflexão crítica do trabalho realizado no âmbito desta tese.

2 Trabalho Relacionado

Foi na primeira edição do HAREM, em 2005, que foi feita pela primeira vez uma avaliação conjunta do reconhecimento de ET em Português. Nesta edição, propôs-se que a categoria TEMPO fosse constituída com os seguintes subtipos:

- Subtipo Data - que englobava todas as referências a dia, mês e ano.
- Subtipo Hora - que englobava todas as referências a horas.
- Subtipo Período - entidade mencionada que referia um intervalo de tempo contínuo e não repetido, com apenas um início e um fim.
- Subtipo Cíclico - entidade que referia períodos recorrentes (véspera de Natal, Páscoa).

Participaram nesta avaliação conjunta 10 sistemas. Na tarefa de identificação do TEMPO e na tarefa de identificação semântica, o sistema que apresentou melhores resultados foi o sistema PALAVRAS-NER.

Mais tarde, em 2008, na segunda edição desta avaliação conjunta, foi novamente feito o reconhecimento de ET. Dos sistemas que participaram no Primeiro HAREM, apenas um participou na segunda edição desta avaliação, o sistema CaGE.

De seguida, descrevem-se sumariamente os sistemas que participaram na categoria de TEMPO no Segundo HAREM.

2.1 Sistemas Participantes no Segundo HAREM

Dos 10 sistemas participantes no Segundo HAREM apenas 3 não realizaram o reconhecimento de expressões temporais. No entanto importa referir também que, os vários sistemas participantes, nesta tarefa se envolveram nela de forma diferenciada.

Sistema	Identificação	TIPO	SUBTIPO	SENTIDO	TEMPO_REF	Normalização
CaGE	X					
PorTexTO	X	X	X			
Priberam	X	X	X		X	
REMBRANDT	X	X	X			
REMMA	X	X				
SeRELep	X	X	X			
XIP-L2F	X	X	X	X	X	X

Tabela 2.1: Sistemas Participantes na Categoria do TEMPO

Na Tabela 2.1 descrevem-se os atributos da categoria do TEMPO que foram preenchidos pelos sistemas participantes nesta tarefa específica de REM. A maioria dos sistemas apenas preencheu o TIPO e SUBTIPO das expressões temporais. apenas um dos sistemas, o XIP-L2F, preencheu todos os atributos, tendo sido o único a preencher os atributos referentes à normalização.

2.1.1 PorTexTo

O sistema PorTexto (Craveiro, Macedo, & Madeira 2008) é um sistema de reconhecimento de entidades temporais em textos de Língua Portuguesa.

Este sistema permite o processamento de documentos em formato de texto simples ou formato estruturado em XML. O resultado produzido pelo sistema pode ser um ficheiro no formato original com as expressões temporais encontradas devidamente anotadas ou um ficheiro com todas as expressões temporais encontradas e a sua posição relativa no texto original.

A identificação de expressões temporais é feita usando padrões de expressões, que são criados a partir de co-ocorrências existentes em referências temporais. O módulo de processamento de co-ocorrências apenas é executado quando não existem padrões de expressões temporais ou quando estes são insuficientes. Este módulo procura identificar as expressões temporais mais utilizadas numa determinada colecção, segundo uma abordagem estatística e procura criar padrões com essas mesmas expressões. Os padrões são armazenados no ficheiro REGEX, facilitando assim a adição de novos padrões e a alteração dos já existentes.

O módulo anotador do Sistema PorTexTO, representado na Figura 2.1 retirada de (Craveiro, Macedo, & Madeira 2008), tem como função identificar as expressões temporais, mediante os padrões de expressões, fazer a sua classificação e, posteriormente, efectuar a anotação do texto.

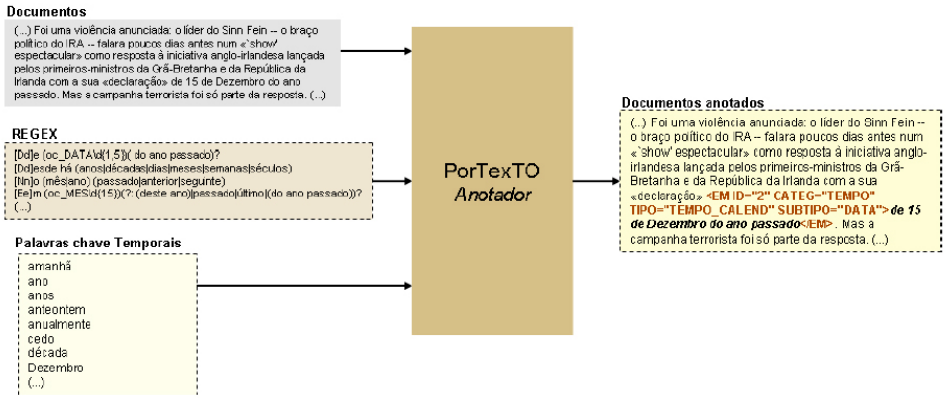


Figura 2.1: Arquitectura do Módulo Anotador do Sistema PorTexTo.

Este módulo tem como entradas o texto original, os padrões de expressões anteriormente criados no módulo Processador de co-ocorrências e uma lista de palavras-chave temporais.

O processamento de um documento de texto é feito frase a frase, sendo cada frase submetida às quatro etapas de processamento deste módulo. Na primeira etapa, decide-se se a frase irá ser processada ou não. Uma lista de palavras-chave é usada para excluir do processamento frases que não possuam expressões temporais. Posteriormente, na segunda etapa, são gerados candidatos a expressões temporais. Depois de identificadas as expressões, estas são marcadas e passam a ser expressões candidatas. A terceira etapa do processamento verifica se existe correspondência entre a frase obtida na etapa anterior e os padrões de expressões definidos anteriormente. Caso exista correspondência, é feita a anotação da expressão e a sua classificação. A última etapa substitui as marcas colocadas durante a segunda fase do processamento pelo texto original.

2.1.2 Sistema de REM da Priberam

O Sistema REM da Priberam (Amaral, Figueira, Mendes, Mendes, Pinto, & Veiga 2008) tem como base um léxico com classificação morfossintática e semântica. A cada entrada no léxico corresponde um ou mais níveis de uma ontologia multilingue que se encontra estruturada através de relações de proximidade conceptual, podendo a cada entrada corresponder um ou mais sentidos, com diferentes valores morfológicos e sintácticos.

Numa primeira fase, a identificação de EM passa pela herança simples dos valores semânticos e morfológicos estabelecidos no léxico. Para contemplar o contexto em que a EM está inserida, é feita a análise contextual sintáctico-semântica.

Este sistema usa regras contextuais que permitem a atribuição ou alteração de valores morfológicos e semânticos a sequências de unidades ou a unidades isoladas. Assim, podem ser criadas locuções através da combinação estrita de sequências de palavras, categorias gramaticais com palavras e combinações de listas de palavras, designadas por "constantes".

As "constantes" têm um papel importante na deteção e classificação de EM, permitindo agrupar palavras com preposições ou outras palavras gramaticais que fazem repetidamente parte de EM. Permitem também detectar e classificar EM através da aglomeração paradigmática de palavras com determinadas afinidades semânticas e morfológicas. Além de listas de palavras e lemas, as "constantes" podem também conter categorias com ou sem restrições morfológicas e semânticas, que podem ser usadas repetidamente nas regras de deteção de EM.

As regras de REM têm em conta as sequências de nomes próprios, separadas ou não por determinadas preposições, assim como o contexto onde são detectadas.

2.1.3 REMBRANDT

O REMBRANDT (Cardoso 2008) é um sistema de reconhecimento de entidades mencionadas e de deteção de relações entre entidades para textos escritos em Português. Este sistema utiliza a Wikipédia como fonte de conhecimento e aplica um conjunto de regras gramaticais que utilizam indícios internos e externos das EM para explicitar o seu significado.

O REMBRANDT possui uma interface própria para interagir com a Wikipédia[1], o SASKIA. Esta interface é responsável por pré-processar as coleções da Wikipédia e realizar uma classificação inicial das EM, com base na informação extraída anteriormente.

As regras gramaticais representam padrões nas frases que indicam a presença de EM com determinadas propriedades semânticas e definem as acções a tomar quando estas são aplicadas com sucesso. As regras são constituídas por uma ou mais cláusulas (unidades de padrões mais simples).

Os documentos são sucessivamente anotados até à sua versão final e é mantido o historial de alterações desde que as expressões são detectadas pela primeira vez até à sua modificação, permitindo assim a afinação do sistema e de regras a aplicar a cada EM.

[1] http://pt.wikipedia.org

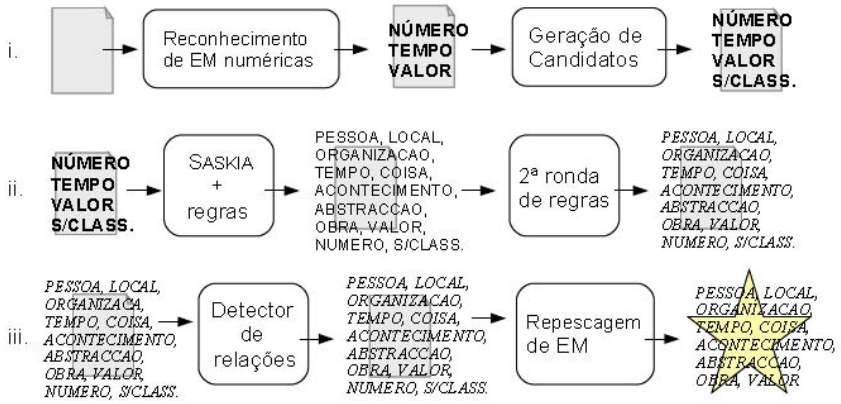


Figura 2.2: Funcionamento do REMBRANT dividido em três etapas.

Na Figura 2.2, retirada de (Cardoso 2008), pode-se observar que o funcionamento do REMBRANDT está dividido em três etapas principais:

- Etapa 1: Reconhecimento de Expressões Numéricas e Geração de Candidatas a EM.

  O Atomizador da Linguateca divide o texto em frases e unidades. Nesta etapa, o primeiro conjunto de regras identifica as expressões numéricas presentes no texto e, posteriormente, são aplicadas regras para a identificação de expressões temporais e de valores. São geradas candidatas a EM pela identificação de sequências de unidades que contêm pelo menos uma letra maiúscula e/ou um algarismo.

- Etapa 2: Classificação de EM.

  Cada uma das expressões candidatas a EM é classificada pelo SASKIA de acordo com os vários significados que a EM pode ter e que são reunidos nas páginas de desambiguação da Wikipédia. Posteriormente, cada expressão candidata é novamente classificada pelo conjunto das regras gramaticais, que englobam indícios externos e internos das EM. Para finalizar a segunda etapa, realiza uma segunda ronda de regras gramaticais utilizando as classificações existentes para detectar EM com morfologias mais elaboradas.

- Etapa 3: Repescagem de EM sem classificação.

  Nesta fase do processo, são identificadas as relações entre EM através de um conjunto de regras específicas para a tarefa. As relações detectadas são utilizadas para repescar algumas EM sem classificação que estão relacionadas com EM anteriormente classificadas.

Figura 2.3: Arquitetura do Sistema REMMA.

Realiza-se uma repescagem de EM com nomes de pessoas, através da comparação com uma lista de nomes comuns. As EM que persistem sem classificação são eliminadas.

2.1.4 REMMA

O sistema REMMA (Ferreira, Teixeira, & Cunha 2008) é um reconhecedor de entidades mencionadas que também utiliza a Wikipédia como fonte de conhecimento externo, em particular através da extração de categorias semânticas da primeira frase das páginas da Wikipédia.

Este sistema encontra-se integrado na Plataforma UIMA (Ferrucci & Lally 2004) (*'Unstructured Information Management Architecture*), uma plataforma livre, escalável e estendível, que possibilita a criação, integração e desenvolvimento de Sistemas de Gestão Estruturada. A plataforma UIMA disponibiliza ferramentas de pré-processamento (leitores, finalizadores genéricos, atomatizadores, separadores em frases e anotadores) e uniformiza a estrutura dos resultados. O UIMA utiliza uma Estrutura de Análise Comum (*Common Analysis Structure*, CAS), o que possibilita aos anotadores o acesso de leitura ao objeto e o acesso de leitura/escrita aos resultados da análise ou às anotações associadas às diferentes regiões dos objetos.

O sistema começa por ler os documentos e guardar os respetivos metadados. Posteriormente, os textos são divididos em frases e átomos, recorrendo às ferramentas de pré-processamento disponibilizadas pela UIMA. O analisador TreeTagger (Schmid 1995) é utilizado na obtenção das categorias morfossintáticas, exclusivamente para eliminar algumas preposições e advérbios candidatos a EM. As anotações geradas são armazenadas nas CAS e posteriormente utilizadas pelos diferentes anotadores do módulo de REM.

O primeiro anotador a ser invocado, o Anotador de Candidatos, identifica todas as expressões candidatas a EM.

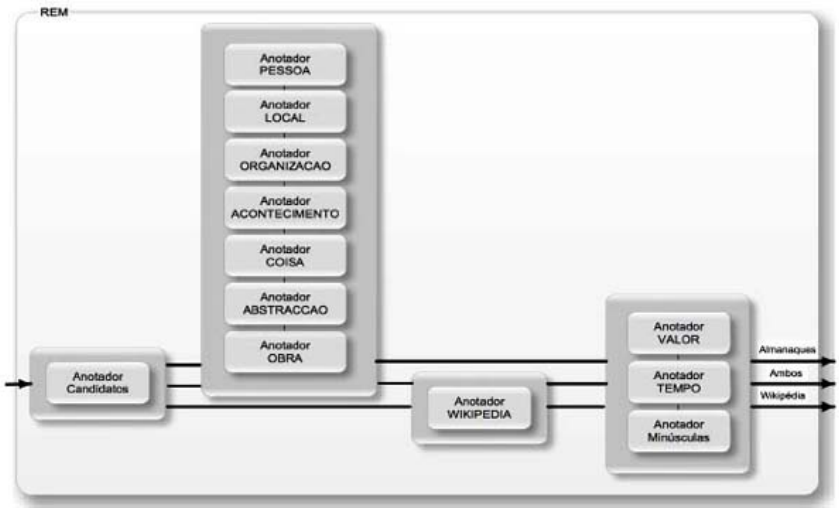


Figura 2.4: Anotadores do Sistema REMMA.

O REMMA contempla duas abordagens de classificação distintas:

- Classificação baseada em Almanaques e regras de análise de contexto.

  Os diversos anotadores dividem a expressão candidata nos seus diversos termos e atribuem uma categoria semântica caso algum dos termos da expressão exista nas listas utilizadas. Se a anotação não é conseguida, os anotadores procuram na expressão candidata palavras pertencentes à classe semântica em análise.

- Classificação com base na informação extraída da Wikipédia.

  Esta classificação pode ser utilizada apenas para entidades candidatas não anotadas anteriormente ou para classificar todas as entidades identificadas. Cada uma das entidades candidatas é convertida num identificador da Wikipédia, o artigo correspondente é recuperado e extrai-se a categoria semântica da entidade em análise a partir da primeira frase do artigo.

Os anotadores "valor" e "tempo" são independentes dos descritos anteriormente. Estes anotadores identificam conjuntos de termos contendo, pelo menos, um algarismo ou pertencentes a uma lista de palavras pré-definidas. Estes anotadores também possuem expressões regulares. O módulo finalizador analisa o CAS e cria os documentos de saída.

2.1.5 CaGe

O Cage (Martins 2008) é um sistema híbrido apoiado por dicionários e regras de desambiguação. Este sistema utiliza dicionários para o reconhecimento das EM das categorias `PESSOA`, `ORGANIZAÇÃO`, `LOCAL` e algumas entidades da categoria `TEMPO`. Para a desambiguação das EM que representam locais ou períodos de tempo, é utilizado um almanaque mais específico, o DIGMAP. O processamento deste sistema pode ser dividido em 4 etapas:

- Etapa 1: Identificação Inicial das Entidades Mencionadas.

  Nesta etapa, os textos são atomizados. Os átomos que anteriormente foram identificados são percorridos de forma sequencial com um algoritmo de modo a extrair o conjunto de sequência de palavras que ocorrem no texto, sendo apenas identificadas sequências de palavras que contêm um máximo de seis elementos (são descartadas as entidades mencionadas no texto com um comprimento superior). Posteriormente é feita uma filtragem e uma pesquisa nos dicionários do sistema para mapear as sequências de palavras com as entidades correspondentes.

  Caso existam vários mapeamentos para uma dada sequência e para as suas subsequências, apenas é considerado o mapeamento que corresponde à sequência de maior tamanho. Para entidades da categoria `LOCAL`, é utilizado um dicionário de excepções para um mapeamento adicional. Se ocorrer um caso de excepção, é retirado o mapeamento entre a sequência de palavra e a entidade.

  Adicionalmente, são utilizadas expressões regulares para identificar entidades da categoria `TEMPO` que não se encontrem nos dicionários.

- Etapa 2 - Classificação das entidades mencionadas e tratamento da ambiguidade.

  Utilizam-se regras de classificação para encontrar a categoria das entidades que foram identificadas nos dicionários e para as quais foram registados vários mapeamentos possíveis. Posteriormente para entidades que continuam ambíguas, faz-se uma escolha circular (round-robin) entre as várias categorias e tipos possíveis.

- Etapa 3 - Desambiguação completa de entidades geográficas e temporais.

Para entidades da categoria `LOCAL` identificadas na etapa anterior, faz-se uma pesquisa no almanaque DIGMAP com o objetivo de associar as entidades aos conceitos geográficos. Se a pesquisa no almanaque retomar vários conceitos geográficos estes são ordenados de acordo com uma heurística do tipo "um sentido por omissão" (Martins, Manguinhas, & Borbinha 2008). Posteriormente, caso as entidades permaneçam ambíguas, é utilizada uma heurística para melhorar a ordenação dos conceitos geográficos correspondentes. Para entidades correspondentes a períodos temporais, faz-se uma pesquisa no almanaque DIGMAP de forma a associar uma entidade ao conceito temporal que lhe está subjacente.

- Etapa 4 - Atribuição de âmbitos geográficos e temporais aos documentos.

  Atribui-se um âmbito geográfico à totalidade do documento, tendo em conta a combinação de todas as referências geográficas identificadas no texto. É atribuído também um âmbito temporal com base no intervalo mínimo que cobre todas as referências temporais identificadas.

2.1.6 SeRELeP

O SeRELeP (Bruckschen, Guilherme Camargo de Souza, Vieira, & Rigo 2008) é um sistema de reconhecimento de relações para textos em Língua Portuguesa. Este sistema foi desenvolvido com o objetivo de participar na pista de reconhecimento de relações do Segundo HAREM. O SeRELeP identifica relações entre EM, sendo as subtarefas de identificação e classificação de EM realizadas pelo analisador sintático PALAVRAS (Bick 2000).

Na Figura 2.5 retirada de (Bruckschen, Guilherme Camargo de Souza, Vieira, & Rigo 2008), encontra-se representado o processo de anotação automática de EM e relações. A entrada do sistema é um texto em formato XML.

As SeRELeP tools são um conjunto de programas auxiliares que são utilizadas no pré-processamento. Estas ferramentas produzem dois corpora: um corpus no formato de texto plano que é utilizado como entrada para o PALAVRAS, e um corpus no formato XML do HAREM, que é a entrada do componente SeRELeP. O componente SeRELeP, para além do XML do HAREM vai necessitar de um corpus anotado pelo PALAVRAS no formato XCES[2].

[2]XML CES: Corpus Encoding Standard for XML, conforme http://www.xces.org/

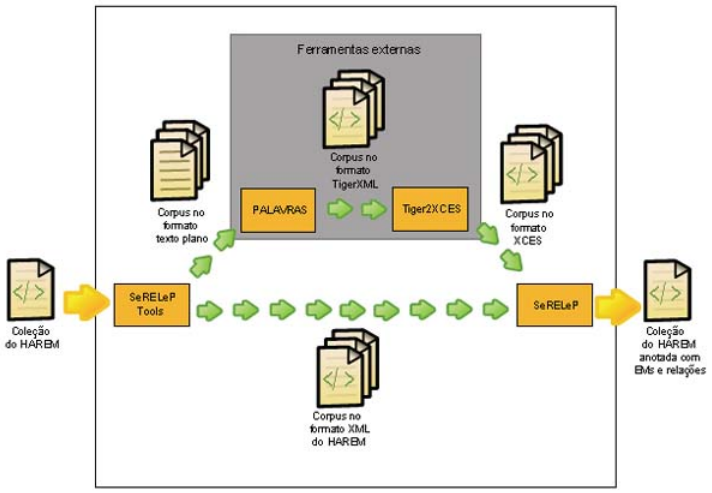


Figura 2.5: Processo de anotação automática de EM e relações do HAREM do sistema SeRELeP.

A anotação do PALAVRAS é codificada em três arquivos XCES: `token`, `pos`, `phrase`, representando cada um deles um nível de anotação linguística:

- O arquivo `token` identifica as unidades lexicais;
- O arquivo `pos` representa informação de nível morfossintático, podendo as etiquetas semânticas estar também representadas neste arquivo;
- O arquivo `phrase` descreve a informação a nível sintático;

Posteriormente o componente SeRELeP, com os dois corpora de entrada, produz um texto anotado com EM e relações. A deteção de relações entre EM é feita com o recurso a heurísticas específicas para cada relação executadas de forma sequencial.

Como dissemos na Introdução, participou ainda nesta edição do HAREM o sistema XIP-L2F em cujo desenvolvimento se integra no trabalho desta dissertação. Por essa razão, apresentá-lo-emos mais adiante a fim de melhor enquadrar o trabalho aqui desenvolvido.

Na secção seguinte, são apresentados os resultados obtidos nas diversas formas de avaliação do TEMPO pelos sistemas anteriormente descritos.

2.2 Resultados da Avaliação no Segundo HAREM

A avaliação da categoria do TEMPO no Segundo HAREM foi feita de forma integrada no HAREM Clássico, ou seja, o tempo foi avaliado juntamente com as outras categorias (cenário total) permitindo que se fizesse uma avaliação global da tarefa de reconhecimento de EM. Adicionalmente, a avaliação tempo foi realizada utilizando a pista do TEMPO (cenário onde os sistemas são avaliados apenas tendo em conta a categoria TEMPO).

A anotação de expressões temporais na Colecção Dourada fez-se em duas fases. Na primeira fase, foram anotadas todas as entidades presentes (incluindo o TEMPO) com os atributos do HAREM Clássico (CATEG, TIPO e SUBTIPO). A CD é constituída por 129 documentos, nas quais estão presentes 1195 entidades de TEMPO.

Posteriormente, numa segunda fase de anotação, foi seleccionado um subconjunto de documentos da CD e as entidades que anteriormente foram classificadas como TEMPO foram anotadas com os restantes atributos (TEMPO_REF, SENTIDO, VAL_NORM e VAL_DELTA). A este subconjunto de documentos chama-se "CD do TEMPO", sendo este composto por 30 documentos com 1508 entidades, das quais 232 são entidades anotadas como TEMPO.

Os sete participantes no Segundo HAREM que adoptaram um cenário que incluía o TEMPO foram avaliados de 4 formas[3]:

- TEMPO Clássico - modo de avaliação de pista do TEMPO que apenas considera os atributos CATEG, TIPO e SUBTIPO. Para este modo de avaliação, utilizou-se a totalidade da CD (ver Tabela 2.2) e a CD do TEMPO (ver Tabela 2.3). Os sistemas que apresentaram melhores resultados com a CD foram os sistemas XIP-L2F, com uma Medida-F de 0.7054 e o sistema PorTexTO, com uma medida de 0,5990. Na CD do TEMPO, os sistemas XIP-L2F e PorTexTO foram novamente os sistemas que obtiveram melhores resultados, com Medida-F de 0,7477 e 0,6177, respectivamente.

  Comparando os resultados obtidos pelos sistemas na classificação do TEMPO com a CD e a CD do TEMPO, verifica-se que os sistemas apresentaram melhores resultados com a CD do TEMPO, o que se pode justificar com o facto da CD e a CD do TEMPO possuírem carateristicas

[3]Cada sistema pôde efectuar 4 corridas, mas neste documento apenas se considerou a melhor corrida de cada um dos sistemas para cada cenário de avaliação.

Tabela 2.2: Classificação do TEMPO - HAREM TEMPO clássico na CD

Sistema	Precisão	Abrangência	Medida-F
XIP-L2F	**0,6812**	**0,7314**	**0,7054**
PorTexTO	0,6694	0,5419	0,5990
REMBRANDT	0,5904	0,4030	0,4790
REMMA	0,4744	0,2538	0,3307
Priberam	0,0832	0,1826	0,1143
Cage	0,0823	0,0294	0,0434
SeRELep	0,0006	0,0016	0,0009

Tabela 2.3: Classificação do TEMPO - HAREM TEMPO Clássico na CD do TEMPO

Sistema	Precisão	Abrangência	Medida-F
XIP-L2F	**0,7376**	**0,7580**	**0,7477**
PorTexTO	0,7350	0,5327	0,6177
REMBRANDT	0,6028	0,4481	0,5140
REMMA	0,4565	0,2581	0,3297
Priberam	0,0950	0,1946	0,1277
Cage	0,0667	0,0166	0,0266
SeRELep	0,0011	0,0028	0,0016

diferentes. Em ambas as CD, as entidades de TEMPO correspondem a 15% das entidades identificadas, no entanto, a distribuição das diferentes categorias de entidades é diferente.

- Tempo Estendido Completo - modo de avaliação da pista do TEMPO, em que todos os atributos de TEMPO são avaliados, incluindo os atributos específicos do TEMPO.

  O sistema que apresentou os melhores resultados (ver Tabela 2.4) foi o sistema XIP, com uma Medida-F de 0,7518. O segundo melhor sistema apresentou uma Medida-F com um valor bastante inferior ao valor obtido pelo melhor sistema. Os resultados baixos obtidos na abrangência pela maioria dos sistemas devem-se provavelmente ao seu diferente grau de envolvimento na tarefa do TEMPO. Tal como foi anteriormente descrito, apenas um sistema participou com todos os atributos. O atributo TEMPO_REF apenas foi atribuído por dois sistemas e o atributo SENTIDO apenas por um sistema.

- Tempo Estendido sem Normalização - modo de avaliação da pista do TEMPO, em que todos os atributos de TEMPO são avaliados com excepção dos atributos de normalização.

Tabela 2.4: Classificação do TEMPO - TEMPO Estendido Completo na CD do Tempo

Sistema	Precisão	Abrangência	Medida-F
XIP-L2F	**0,8087**	**0,7024**	**0,7518**
PorTexTO	0,7350	0,2708	0,3957
REMBRANDT	0,6028	0,2277	0,3306
REMMA	0,4565	0,1313	0,2038
Priberam	0,1115	0,1186	0,1150
Cage	0,0667	0,0166	0,0266
SeRELep	0,0011	0,0014	0,0012

Tabela 2.5: Classificação do TEMPO - Estendido sem Normalização na CD do TEMPO

Sistema	Precisão	Abrangência	Medida-F
XIP-L2F	**0,7733**	**0,7472**	**0,7600**
PorTexTO	0,7350	0,3730	0,4949
REMBRANDT	0,6028	0,3137	0,4127
REMMA	0,4565	0,1807	0,2589
Priberam	0,1115	0,1634	0,1326
Cage	0,0667	0,0116	0,0198
SeRELep	0,0010	0,0019	0,0014

Assim, são avaliados os atributos CATEGORIA, TIPO, SUBTIPO, TEMPO_REF e SENTIDO. (ver Tabela 2.5)

- Tempo Estendido só com Normalização - modo de avaliação da pista do TEMPO em que são avaliados os atributos TEMPO previstos no HAREM clássico e também os atributos de normalização. São avaliados os atributos CATEGORIA, TIPO, SUBTIPO, VAL_NORM e VAL_DELTA. (ver Tabela 2.6)

A categoria TEMPO no Segundo HAREM foi avaliada de forma flexível, permitindo observar, de forma independente, o comportamento global dos sistemas no reconhecimento e classificação das entidades de TEMPO e seu comportamento na normalização.

Nos quatro modos em que foram avaliados os sistemas, há grande disparidade entre os sistemas, no entanto, os sistemas mantiveram a sua ordenação relativa. Os sistemas que apresentaram melhores resultados foram os sistemas XIP-L2F e o PorTexTO. Na avaliação dos atributos do TEMPO no HAREM Clássico, a diferença entre as duas Medidas-F é aproximadamente 0,1. No

Tabela 2.6: Classificação do TEMPO - TEMPO Estendido com Normalização na CD do TEMPO

Sistema	Precisão	Abrangência	Medida-F
XIP-L2F	**0,7908**	**0,6970**	**0,7410**
PorTexTO	0,7350	0,3461	0,4706
REMBRANDT	0,6028	0,2911	0,3926
REMMA	0,4565	0,1676	0,2452
Priberam	0,0950	0,1264	0,1085
Cage2	0,0667	0,0108	0,0186
SeRELep	0,0011	0,0018	0,0014

entanto, na avaliação do TEMPO Completo, a diferença é significativamente maior: 0,35.

De seguida são descritas as directivas adoptadas no Segundo HAREM. A principal motivação de apresentar aqui a metodologia seguida na identificação, classificação e normalização do TEMPO, é permitir uma comparação com as diretivas seguidas neste documento. No Capítulo 3 são apresentadas as principais diferenças entre os dois conjuntos de directivas.

2.3 Directivas do TEMPO no Segundo HAREM

A proposta de anotação da categoria do TEMPO (Hagège, Baptista, & Mamede 2008) para o Segundo HAREM, surge como seguimento à proposta de (Cardoso & Santos 2007), tendo sido enriquecida e alargada de forma a abranger a uma noção mais lata de expressão temporal. As diretivas contemplam ainda a possibilidade de normalizar EM do tipo TEMPO tendo em vista o cálculo de referências temporais.

De seguida, descreve-se o conjunto de diretivas elaboradas para o Segundo HAREN tendo como objectivo o reconhecimento, classificação e normalização de ET.

A Categoria TEMPO engloba expressões temporais que, semanticamente, denotam um momento no calendário (ponto ou intervalo) e expressões de quantificação temporal que exprimem uma duração ou uma repetição de eventos. Esta categoria também engloba o emprego genérico de algumas expressões associadas à noção de tempo.

No Segundo HAREM, foram definidos os critérios seguintes[4] para a identificação de ex-

[4]Os critérios descritos podem ser consultados na íntegra em (Hagège, Baptista, & Mamede 2008)

pressões temporais. Note-se que são consideradas ET, as expressões que correspondem ao critério 1 e adicionalmente a pelo menos um dos subcritérios de 2 ou ainda ao critério 3.

- **Critério 1 -** uma expressão temporal em contexto pode responder adequadamente a uma das interrogativas *"(<prep>) quando?"*, *"(<prep>) quanto tempo?"* , *"(<haver>) quanto tempo?"* ou *"com que frequência?"*.

- **Critério 2 -** uma expressão temporal contém pelo menos uma unidade lexical que corresponda a um dos seguintes tipos:

  1. uma data numérica (por exemplo, 29-10-2008);
  2. uma unidade de medida temporal (*dia, mês, trimestre, ano, século, etc.*) ou um advérbio terminado em *"-mente"* derivado destas expressões (*diariamente, semanalmente, mensalmente,* etc.);
  3. um nome correspondente à designação de uma destas unidades de medida de tempo, isto é, nome de mês (*Setembro, Dezembro,* etc.), nome de dia (*segunda-feira, domingo,* etc.);
  4. um nome de festividade, religiosa ou não (*Natal, Páscoa, Quaresma, Entrudo*); nomes de estações do ano (*Primavera, Inverno*); nomes de festividades, que podem incluir o nome *dia (dia de Santo António, dia de Nossa Senhora da Conceição, dia de São Valentim, dia dos namorados, no São Martinho,* etc.);
  5. alguns advérbios de tempo (simples, não derivados e semanticamente não ambíguos), tais como: *hoje, ontem, amanhã, outrora;* algumas locuções adverbiais de tempo (ou advérbios compostos), semanticamente não ambíguas, como, por exemplo: *antes de ontem, depois de amanhã;* excluiram da actual campanha de avaliação os seguintes advérbios simples, sintáctica ou semanticamente ambíguos, apesar de poderem representar expressões temporais: *agora, ainda, já, sempre.*
  6. um sintagma preposicional cujo núcleo seja uma das palavras *altura, tempo, momento, período, era,* etc., quando estas palavras forem determinadas por um demonstrativo (por exemplo:*nesse tempo)*, ou especificados por uma relativa (por exemplo: *na altura em que ela adoeceu*), um possessivo (por exemplo: *durante a nossa era*) ou modificado por outro sintagma preposicional introduzido por de (por exemplo: *durante a era dos*

dinossauros) ou então por um adjectivo capitalizado (por exemplo:*durante o período Barroco, Cretáceo,* etc.);

7. Os complementos determinativos com a forma de *Num Ntmp* de nomes predicativos, que não respondem adequadamente ao critério (1) mas que são indubitavelmente EM a anotar *(e.g. uma viagem de 5 dias)*; a preposição *de* deve ser incluída na EM;
8. expressões de frequência, como as seguintes: *de vez em quando, às vezes, de quando em quando, frequentemente,* etc.
9. expressões da forma Prep + <unidade de medida temporal> +*que* + verbo *vir* ou verbo *passar* (por exemplo, *no ano que passou*, *para o mês que vem*)
10. expressões com os verbos *fazer* ou *haver* e <unidade de medida temporal> (e.g. *há três anos, faz duas semanas*).

Notas:

- Excluem-se, no Critério 1, as expressões de tipo genérico como o emprego de o *Inverno* em frases como *Adoro o Inverno*, que serão retomadas de forma autónoma no critério 3, abaixo.
- Excluem-se também, com o Critério 1, as locuções que, embora contendo expressões do conjunto identificado no critério 2, não respondem adequadamente às interrogativas de tempo. Trata-se de locuções como *de dia para dia*, que se encontram em frases como *A situação agrava-se de dia para dia*, que funciona como um circunstancial de modo; repare-se na inaceitabilidade do par pergunta-resposta: P:*(quando, com que frequência, em quanto tempo) é que a situação se agrava?* R: *de dia para dia.*
- Repare-se também que, para qualquer dos pontos 2-2 a 2-9 do critério 2, se pode fazer uma definição em *extensão* dos elementos em questão. Assegura-se, assim, o problema de intersubjetividade das anotações.

- **Critério 3 -** uma expressão temporal que contém uma unidade lexical do tipo das que foram definidas no critério 2 mas para a qual o critério 1 não se aplica. Trata-se de expressões temporais genéricas como o *mês de Julho* em exemplos como *Adoro o mês de Julho* onde o *mês de Julho* não responde à pergunta *quando?* embora contenha elementos lexicais como os que foram definidos no critério 2.

2.3.1 Delimitação de Expressões Temporais no Segundo HAREM

No HAREM, foi definido que a totalidade da expressão temporal deverá ser delimitada por <EM ID=... CATEG="TEMPO"> e </EM>, incluindo a preposição que a introduz, no caso de a expressão temporal ser um sintagma preposicional, ou um determinante, no caso de ser um sintagma nominal.

Para as expressões temporais complexas, o Segundo Harem adoptou os critérios de segmentação descritos em (Hagège & Tannier 2007). Assim, uma expressão temporal complexa deverá ser dividida em unidades menores se se verificarem os seguintes critérios:

1. cada expressão componente é sintaticamente válida quando combinada independentemente com o evento que modifica.
2. cada expressão componente, combinada com o evento que modifica, está logicamente implicada na expressão complexa, ou seja, cada combinação "evento + expressão_temporal_mínima" deve ser logicamente implicada pela combinação "evento + expressão_temporal_complexa". Por outras palavras, o valor de verdade de todas a combinações "evento + expressão_temporal_mínima" deve poder ser deduzido do valor de verdade da combinação

   "evento + expressão_temporal_complexa".

2.3.2 Tipos e Subtipos de Entidades TEMPO no Segundo HAREM

No HAREM, todas as ET deverão possuir o atributo obrigatório TIPO. Este atributo é o único definido como obrigatório do elemento EM na categoria TEMPO.

Nas subsecções seguintes, os diferentes tipos, subtipos e atributos são descritos com mais detalhe.

2.3.2.1 Tipo TEMPO_CALEND

São consideradas expressões do tipo TEMPO_CALEND expressões que permitem inserir o predicado que modificam numa linha temporal, seja como ponto ou como intervalo. O tipo TEMPO_CALEND possui os seguintes subtipos:

- **Subtipo DATA** O subtipo `DATA` engloba datas absolutas, que contêm informação necessária para uma localização no calendário e possuem os campos `DD-MM-AA` (dia-mês-ano), e por outro lado, datas relativas, para as quais é necessário estabelecer um ponto de referência para as poder localizar na linha temporal.

  Exemplos:

  Data absoluta completa, com os campos dia, mês e ano preenchidos

  Vou viajar <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" >**no dia 19 de Outubro de 2007** </EM>

  Data relativa

  Vou a Lisboa <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA"> **no próximo dia 22** </EM>

- **Subtipo HORA** O subtipo `HORA` engloba expressões temporais com valor de `DATA`, mas com granularidade inferior à unidade dia.

  O Pedro está disponível <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA"> **às 15:00**</EM>

- **Subtipo INTERVALO** Este subtipo engloba expressões que localizam uma ação entre dois limites temporais, pelo que corresponde a uma expressão composta por duas expressões temporais.

  Trabalhei em Londres <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO"> **entre 2000 e 2003**</EM>

2.3.2.2 Tipo DURACAO

O tipo `DURACAO` refere uma duração de tempo contínuo que expressa quantificação temporal e responde à pergunta *"quanto tempo?"*. Não expressa a localização de um evento, mas sim quantificação temporal.

Fiquei em Lisboa <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**dois meses** </EM>

2.3.2.3 Tipo FREQUENCIA

O subtipo FREQUENCIA engloba expressões temporais que exprimem uma repetição ao longo da linha temporal.

Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA"> **diariamente** </EM>

2.3.2.4 Tipo GENERICO

O tipo GENERICO não refere uma data específica, no entanto, é composta por elementos temporais.

Adoro <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">**o Verão.** </EM>

2.3.3 Atributos do TEMPO Presentes no Segundo HAREM

2.3.3.1 Atributo TEMPO_REF

O atributo TEMPO_REF apenas está presente nas expressões temporais do tipo TEMPO_CALEND e subtipo DATA. No caso de datas absolutas, o valor do atributo toma o valor ABSOLUTO. No caso de datas referenciais que fazem referência ao tempo de enunciação, toma o valor ENUNCIACAO. Quando a inferência temporal deve ser considerada no próprio texto, este atributo toma o valor TEXTUAL.

Nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO">**a 3 de Janeiro de 1986.** </EM>

Nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**dois dias depois do Natal** </EM>

Nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO">**na sexta-feira passada** </EM>

2.3.3.2 Atributos SENTIDO e VAL_DELTA

No caso de expressões temporais referenciais, ou seja, quando o valor do atributo TEMPO_REF tem o valor TEXTUAL ou ENUNCIACAO, os atributos SENTIDO e VAL_DELTA podem estar presentes na anotação.

O atributo SENTIDO indica se o valor temporal da expressão se situa cronologicamente antes, depois, ou simultâneo ao tempo de referência. Os possíveis valores para este atributo são: ANTERIOR, POSTERIOR, SIMULT, ANTERIOR_OU_SIMULT, POSTERIOR_OU_SIMULT.

Quanto ao atributo VAL_DELTA, este indica a distância temporal entre o tempo do evento indicado pela expressão temporal e o momento de referência, sempre que esta distância temporal aparece explicitamente no texto. No caso desta distância temporal não ser explícita, o valor de VAL_DELTA é omitido.

O atributo VAL_DELTA possui o seguinte formato:

A<digitos>M<digitos>S<digitos>D<digitos>H<digitos>M<digitos>S<digitos>

Onde:

As letras A, M, S, D, H, M, S são constantes que devem aparecer nesta ordem e que correspondem respectivamente, aos valores de Anos, Meses, Semanas, Dias, Horas, Minutos e Segundos.

Os <digitos> à direita das letras correspondem ao número de Anos, Meses, Semanas, Dias, Horas, Minutos e Segundos que se devem adicionar à data de referência para obter o valor temporal da expressão anotada.

Exemplos:

Veio <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO" SENTIDO="ANTERIOR"VAL_DELTA="A0M0S0D1H0M0S0"> **ontem** </EM>

Nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S0D2H0M0S0">**dois dias depois do Natal**</EM>

2.3.3.3 Atributo VAL_NORM

O atributo `VAL_NORM` reflecte a normalização de expressões temporais e apenas é atribuído a algumas entidades de `TEMPO`. Este atributo, consoante o tipo de ET, sofre alterações no formato.

Para expresssões do subtipo `DATA`, com o atributo `TEMPO_REF` com valor `ABSOLUTO`, o valor de `VAL_NORM` obedece ao seguinte formato:

`<Era><Ano><Mes><Dia>T<Hora><Minuto>E<ESTACAO>LM<limite_aberto>`

Onde:

`<Era>` corresponde a 1 caracter que é + ou - conforme a data seja depois ou antes da nossa era;

`<Ano>` corresponde a 4 caracteres de tipo dígito que representam o valor do ano ou então a subsequência "–";

`<Mes>` corresponde a 2 caracteres de tipo dígito que representam o valor do mês ou então a subsequência "-";

`<Dia>` corresponde a 2 caracteres de tipo dígito que representam o valor do dia ou então a subsequência "-";

`<Hora>` corresponde a 2 caracteres de tipo dígito que representam o valor da hora ou então a subsequência "-";

`<Minuto>` corresponde a 2 caracteres de tipo dígito que representam o valor dos minutos ou então a subsequência "-";

`<ESTACAO>` corresponde a duas letras capitalizadas correspondente às estações do ano. `IN` para Inverno, `PR` para Primavera, `VE` para Verão e `OU` para Outono. No caso da data absoluta não ser expressa em termos de estação do ano, este campo terá por valor a subsequência "_";

`<limite_aberto>` indica se a expressão normalizada de data absoluta introduz um intervalo de tempo com limite anterior ou limite posterior não determinado (em aberto). Os valores respetivos são `''A''` (no caso de limite anterior em aberto; este caso a expressão temporal

corresponde ao limite posterior); ‘‘P’’ no caso de limite posterior em aberto ; “-” quando a data absoluta não corresponde a um intervalo com um dos limites aberto.

Exemplo:

Nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+19860103T--E-LM-"> **a 3 de Janeiro de 1986**</EM>

No caso do subtipo HORA, é utilizado o mesmo formato, no entanto, os campos <Era><Ano><Mes><Dia> correspondem à sequência "+---" e o campo <ESTACAO> toma o valor de -" .

Está disponível <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+---T15-E-LMA"> **antes das 3:00 da tarde**</EM>

Para expressões do tipo DURACAO, o atributo VAL_NORM expressa uma distância temporal e possui o seguinte formato:

A<digitos>M<digitos>S<digitos>D<digitos>H<digitos>M<digitos>S<digitos>

Vivi em Lisboa <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="A0M2S0D0H0M0S0"> **dois meses**</EM>

Na secção seguinte é descrita a cadeia de processamento STRING, a fim de melhor compreender o trabalho realizado no âmbito desta dissertação, cuja implementação será descrita mais adiante, no Capítulo 4.

2.4 Cadeia de Processamento L^{2}F (STRING)

A cadeia de processamento de Língua Natural do L^{2}F é constituída por vários módulos, que se encontram representados na Figura 2.6 imagem retirada de (Mamede 2011). Entre os diferentes módulos da cadeia de processamento é utilizado XML (eXtensible Markup Language). Este trabalho integra-se no desenvolvimento do último módulo de processamento, o XIP.

Figura 2.6: Arquitectura da Cadeia de Processamento do L²F.

O primeiro módulo da cadeia é responsável pela segmentação. Este módulo divide o texto em tokens, identificando endereços de e-mail, números ordinais terminados com um "º"ou com "ª", números com "."e ",", números IP e endereços url, inteiros, abreviaturas com vários "."e sequências de "?", "!"e "...", sinais de pontuação, símbolos, números romanos, sequência de caracteres não aceites na etiquetação morfossintática e palavras. Observe-se como exemplo a segmentação aplicada à frase *"A Maria vai para Lisboa hoje."* (Figura 2.7).

```
word[0]:   |A|
word[1]:   |Maria|
word[2]:   |vai|
word[3]:   |para|
word[4]:   |Lisboa|
word[5]:   |hoje|
word[6]:   |.|
```

Figura 2.7: Segmentação Aplicada a uma frase

A etiquetação morfossintática é realizada pelo sistema Palavroso (Medeiros 1995), sendo atribuídas todas as etiquetas possíveis aos tokens identificados anteriormente. É utilizado um conjunto de etiquetas PAROLE, que contêm informação codificada em 10 campos (categoria,

```
word[0]:   |A|       POS->[a]S....==..s
                          [a]Td...sf...
                          [a]Pp..3sf.as
word[1]:   |Maria|   POS->[Maria]Np...s=...
word[2]:   |vai|     POS->[ir]V.ip3s=...
                          [ir]V.m=2s==..
word[3]:   |para|    POS->[para]S....==..s
word[4]:   |Lisboa|  POS->[Lisboa]Np...s=...
word[5]:   |hoje|    POS->[hoje]R......p..
word[6]:   |.|       POS->[.]O.........
```

Figura 2.8: Etiquetação morfossintática uma frase

subcategoria, mood, tense, pessoa, número, género, grau, caso, formação) e 13 categorias (nome, verbo, adjetivo, pronome, artigo, advérbio, preposição, conjugação, numeral, interjeição, marcador passiva, residual e pontuação). Pode observar-se a aplicação dessas etiquetas à frase anterior.

O módulo seguinte divide o texto em frases, sendo considerados indicadores de fim de frase todos os segmentos constituídos unicamente por "." "!" e "?". Como excepções são consideradas as abreviaturas previamente registadas e a existência de "≫", ")", "]", "" sempre que estas ocorram a seguir a "...)" .

O módulo de desambiguação morfossintático, RudriCo2[5] (Diniz 2010), é uma evolução do sistema Rudrico (Pardal 2007). Este módulo executa correções à saída do módulo de etiquetação morfossintática, efetuando alterações nos lemas, como é o caso da palavra "quaisquer". Este módulo também executa regras de contracção para analisar as contrações (e.g. desta = de + esta) ou para juntar num único token palavras composta (e.g. bom senso).

```
camada> |contexto_esquerdo| antecedente |contexto_direito| :< consequente
```

Refira-se também que o antecedente, o consequente e os contextos são constituídos por itens da forma:

```
[prop_1='valor 1', prop_2='valor 2' ...  ][...]
```

[5]Rule Driven Converter

As regras são organizadas por camadas e são aplicadas de acordo com a camada a que pertencem, começando pela camada de menor número. Dentro da mesma camada, usa-se a ordem pela qual as regras são declaradas no ficheiro.

O antecedente de uma regra define o padrão que tem de existir para que uma regra se aplique. O `contexto_esquerdo` e o `contexto_direito` são opcionais. Quando o padrão especificado pelo antecedente se verifica e os contextos emparelham com os segmentos envolventes, a sequência de segmentos que emparelha com o antecedente é substituída por aquilo que se encontra no consequente.

Assim, por exemplo, atente-se na regra para analisar a contracção "das", em que a forma superficial é substituída pelas formas "de"e "as, a que são associadas as respectivas categorias gramaticais. [6]

```
 [surface='das', lemma='de', CAT='partd', NUM='p', GEN='f']

:<

[surface='de', lemma='de', CAT='prep'],

[surface='as', lemma='o', CAT='artd', NUM='p', GEN='f'] .
```

Adicionalmente, este módulo também realiza regras de desambiguação para escolher qual das anotações está certa e quais devem ser descartadas. Estas regras têm a mesma sintaxe das regras de recomposição, mas o símbolo que separa o antecedente do consequente é o :=.

O módulo de desambiguação morfossintática estatística, o Marv (Ribeiro, Mamede, & Trancoso 2003), efetua uma desambiguação estatística em que seleccionando uma das etiquetas associadas a cada palavra e utilizando para tal o algoritmo de Viterbi (Jurafsky & Martin 2000). Essa escolha é baseada apenas na Categoria e Subcategoria da palavra.

2.4.1 XIP

O XIP é o último módulo da cadeia de processamento e é responsável pela análise sintática. O XIP é um parser que permite introduzir informação lexical, sintática e semântica, e utiliza

[6] Ignora-se neste exemplo a ? entre o artigo e o chamado "pronome demonstrativo."

essa informação para processar as frases, analisando-as em sintagmas nucleares (ou chunks) e calculando dependências entre estes. Neste processo, é ainda possível aplicar gramáticas locais e regras de desambiguação. Estes últimos serão descritos já a seguir.

O XIP é composto por diferentes módulos:

- Léxicos - estes permitem adicionar informação aos diferentes tokens. No XIP, as expressões linguísticas já anotadas, provenientes da ferramenta de etiquetagem morfossintáctica, podem ser agora enriquecidas adicionando entradas lexicais ou alterando os atributos das existentes. Por exemplo, na seguinte regra lexical adiciona-se ao lema `noite` (que é um nome) os traços `time` e `part_of_day`:

  ```
  noite:  noun += [time:+, part_of_day:+].
  ```

- Módulo de Desambiguação - módulo que efectua desambiguação por regras de certas expressões linguísticas. A sintaxe para estas regras é a seguinte:

  ```
  camada> filtro_tokens = |contexto_esquerdo| etiquetas_seleccionadas
                          |contexto_direito|.
  ```

  O `filtro_tokens` corresponde a uma expressão que especifica um subconjunto de categorias ou traços associados a um token, enquanto `etiquetas_seleccionadas` corresponde a categorias ou traços que se pretende selecionar de entre as entradas lexicais encontradas pelo `filtro_tokens`. O `contexto_esquerdo` e o `contexto_direito` são opcionais.

  Observe-se, como exemplo, a palavra "Natal", que pode ser uma localização geográfica (no Brasil) ou o nome da festividade. No entanto, se for antecedida por "antes de" é possível com alguma segurança indicar que se trata de uma festividade. Tal é feito pela seguinte regra:

  ```
  20> noun[maj,surface:Natal] %= | prep[lemma:"antes_de"], (art;?[dem]),
                                 (adj) | noun[one_day=+,proper=+].
  ```

  Nesta regra de desambiguação, selecciona-se todas as entradas lexicais que tenham a categoria `noun` (nome) e os traços `maj` (maiúscula) e `surface` com o valor "Natal". Se o

contexto à esquerda for uma preposição com o lema "antes de", seguido opcionalmente de um artigo ou de um pronome demonstrativo e podendo ainda ser seguido opcionalmente por um adjectivo, atribui-se, então, ao nó os traços `one_day` e `proper`.

- Gramáticas Locais - o XIP possibilita a escrita de regras considerando os contextos à esquerda e à direita. As regras definidas nas gramáticas locais pretendem identificar e classificar expressões constituídas por mais do que uma unidade lexical; ou seja, nas situações em que diferentes unidades lexicais não devem ser consideradas individualmente, as regras juntam esses elementos numa única entidade. No exemplo seguinte, constroí-se um nó `noun` com os traços `time`, `date` e `tipo_tempref` que junta duas unidades lexicais: um `noun` com o lema `período` e outro nó com o lema `paleolítico`.

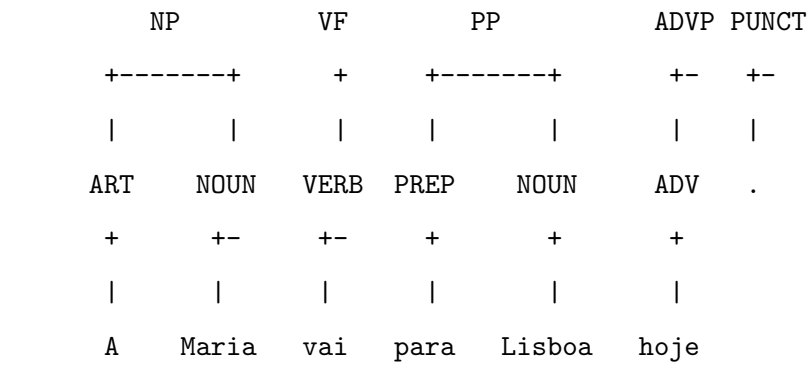
```
2> noun[time=+, date=+, tipo_tempref=absolut] = noun[lemma:período],
                                                ?[lemma:paleolítico].
```

- Módulo de Chunking - este módulo realiza uma análise sintática preliminar do texto, construindo para cada frase uma sequência de sintagmas nucleares (chunks). Na Figura 4, encontra-se representada a árvore obtida pela análise sintática de uma frase, depois de terem sido aplicadas as regras de chunking.

- Módulo de Dependências - as dependências representam relações sintáticas entre os diferentes chunks. A sequência de nós identificados anteriormente pelas regras de chunking é utilizada pelas regras de dependência para calcular as relações entre eles[7]. Na Figura 4, encontram-se representadas as dependências calculadas para a frase que temos vindo a usar como exemplo. Assim, a dependência `SUBJ_PRE(vai,Maria)` indica que `Maria` é sujeito de `vai` , enquanto que o sufixo `_PRE` assinala que o sujeito se encontra à esquerda do verbo.

```
                        TOP
      +----------+---------+------------+----+
      |          |         |            |    |
```

[7]Apenas as dependências principais foram apresentadas na figura, tendo-se omitido um conjunto de dependências auxiliares como, por exemplo, a relação HEAD entre um constituínte e a sua cabeça, e.g., a relação entre um nome e o sintagma nominal de que este é o núcleo




```
DETD(Maria,A)
VDOMAIN(vai,vai)
MOD_POST(vai,hoje)
MOD_POST(vai,Lisboa)
SUBJ_PRE(vai,Maria)
NE_INDIVIDUAL_PEOPLE(Maria)
NE_LOCAL_CITY_ADMIN_AREA(Lisboa)
NE_REF-SIMULT_TREF-ENUNC_TEMPO_DATE(hoje)
0>TOP{NP{A Maria} VF{vai} PP{para Lisboa} ADVP{hoje} .
```

3 Novas Directivas

Na sequência do Segundo HAREM foram apontadas certas limitações às directivas então adoptadas, tendo sido feitas sugestões para trabalho futuro. As directivas aqui propostas surgem no seguimento das directivas do Segundo HAREM, pretendendo-se resolver assim algumas das limitações e ampliar o conjunto de número de ET a classificar e normalizar.

As directivas encontram-se descritas na íntegra em (Baptista, Mamede, Hagège, & Maurício 2009), apresentando-se neste capítulo apenas um resumo das principais modificações, introduzidas às diretivas do Segundo HAREM, que resultam do trabalho desenvolvido nesta dissertação.

TIPO	SUBTIPO
TEMPO_CALEND	DATA
	HORA
	INTERVALO
DURACAO	
FREQUENCIA	
GENERICO	

TIPO	SUBTIPO
TEMPO_CALEND	DATA
	INTERVALO
	COMPLEXO
	RESIDUAL
DURACAO	SIMPLES
	INTERVALO
FREQUENCIA	
GENERICO	

Alterações Introduzidas na Classificação de ET. Directivas do 2º Harem - Proposta Actual

A fim de melhor compreender as alterações introduzidas na classificação, apresenta-se, nas tabelas anteriores, os tipos e subtipos definidos no Segundo HAREM e na presente proposta, respectivamente. Esta proposta mantem o conjunto de tipos e introduz alguns subtipos novos. Nas secções seguintes apresentamos com pormenor as restantes modificações.

3.1 TIPO TEMPO_CALEND

O tipo `TEMPO_CALEND` é redefinido com os seguintes subtipos:

`DATA`	engloba datas absolutas e datas relativas, o subtipo `DATA` inclui agora as expressões temporais com granularidade inferior ao dia (horas). Exemplo: *"O João chegou a Londres no dia 3 de Maio às 21:45."*
`INTERVALO`	expressões compostas por duas `DATAS`. Exemplo: *"O João viveu em Lisboa entre Abril 1995 e Junho de 2000."*
`COMPLEXO`	expressões compostas por uma `DATA` e uma `DURAÇÃO`. Exemplo: *"Faz no Verão de 2008 três anos que o João se mudou."*
`RESIDUAL`	reúne expressões que não podem ser classificadas em nenhum dos subtipos anteriores; este subtipo não possui qualquer outro atributo. Exemplo: *"O João chegou fora de horas."*

3.2 Tipo DURACAO

O tipo `TEMPO_DURACAO` passa a ter os seguintes subtipos:

`SIMPLES`	subtipo que expressa uma quantificação temporal composta pelo menos por um quantificador e uma unidade de medida temporal. Exemplo: *"A reunião durou 1 hora e 20 minutos."*
`INTERVALO`	tipo complexo composto por sequências de sintagmas preposicionais, em que o primeiro termo designa a duração menor e o segundo a duração maior; observam-se os padrões de preposições "de...a","entre...a" e a coordenação "entre...e". Exemplo: *"O João esperou pelo autocarro entre 20 a 30 minutos."*

3.3 Atributo VAL_NORM

O atributo VAL_NORM passa a ter o seguinte formato:

VAL_NORM="<Era>**M**<Milenio><Seculo><Decada>**D**<Ano><Mes><Dia>
T<Hora><Minuto><Segundo><Milissegundo>**E**<Estacao>**LM**<limite_aberto>"

O formato do atributo VAL_NORM passa a permitir a normalização de horas até ao milissegundo e a normalização de milénios, séculos e décadas. Este atributo apenas está presente no tipo TEMPO_CALEND. Dependendo do subtipo a que se aplica, o seu formato geral pode sofrer adaptações:

- Para TEMPO_CALEND e subtipo DATA ou subtipo COMPLEXO o formato mantém-se. No entanto, as ET do subtipo COMPLEXO, que envolvem uma DATA e uma DURACAO, apenas podem ser normalizadas se a DATA for absoluta.

O João foi ao ginásio <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+CM--------D20080105T---------E--LM-"SENTIDO="SIMULT"> **no dia 5 de Janeiro de 2008** </EM>

- Para TEMPO_CALEND e subtipo INTERVALO, o atributo VAL_NORM é duplicado e são atribuídos índices numéricos (VAL_NORM1 e VAL_NORM2), cada um correspondendo a cada uma das datas que limitam o intervalo.

O ginásio está encerrado <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ABSOLUTO" TEMPO_REF2="ABSOLUTO" VAL_NORM1="+CM--------D20080105T---------E--LM-" VAL_NORM2="+CM--------D20080115T---------E--LM-"> **entre 5 e 15 de Janeiro de 2008**</EM>

3.4 Atributo VAL_DELTA

O atributo VAL_DELTA possui o seguinte formato :

VAL_DELTA="**A**<**digitos**>**D**<**digitos**>**H**<**digitos**>**M**<**digitos**>**S**<**digitos**>**M**<**digitos**>"

Onde:

As letras `A, D, H, M, S e M` são constantes que devem aparecer por esta ordem e que correspondem respectivamente, aos valores de Anos, Meses, Semanas, Dias, Horas, Minutos Segundos e Milissegundos.

Os <digitos> à direita das letras constantes correspondem ao valor numérico de Anos, Dias, Horas, Minutos, Segundos e Milissegundos que se devem adicionar (ou subtrair) à data de referência para calcular o valor temporal da expressão anotada. O atributo VAL_DELTA também é utilizado para expressar durações.

A alteração do formato deste atributo, veio permitir normalizar durações inferiores a segundos. O formato de VAL_DELTA pode variar consoante o tipo e subtipo da ET:

- Para expressões do tipo TEMPO_CALEND referenciais, o formato é o descrito acima. O valor de VAL_DELTA indica a distância temporal entre o tempo do evento a que está associada pela expressão temporal e o momento de referência. O valor <digit> pode ser negativo, quando distância temporal deve ser subtraída ao momento de referência

Isso aconteceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO"VAL_DELTA="A0D-1H0M0S0M0"SENTIDO="SIMULT">**ontem**</EM>

- Para TEMPO_CALEND e subtipo INTERVALO com datas relativas, o atributo VAL_DELTA é duplicado e são atribuídos índices numéricos (VAL_DELTA1 e VAL_DELTA2), cada um correspondendo a cada uma das datas que limitam o intervalo.

- Para DURACAO e subtipo SIMPLES, o formato é o descrito acima.

Ele viveu em Lisboa <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" SUBTIPO="SIMPLES" VAL_DELTA="A2D0H0M0S0M0">**2 anos**</EM>

- Para DURACAO e subtipo INTERVALO, o atributo VAL_DELTA mantém o seu formato, no entanto, é duplicado (VAL_DELTA1 e VAL_DELTA2), cada um correspondendo a cada um dos valores da duração.

A reunião durou <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" SUBTIPO="INTERVALO" VAL_DELTA1="A0D0H2M0S0M0" VAL_DELTA2="A0D0H3M0S0M0"> **entre 2 a 3 horas**</EM>

- Para o tipo FREQUENCIA este atributo passa a apresentar o formato:

  VAL_QUANT="<digitos>"
  VAL_DELTA="**A**<**digitos**>**D**<**digitos**>**H**<**digitos**>**M**<**digitos**>**S**<**digitos**>**M**<**digitos**>"

  O VAL_DELTA indica a granularidade da expressão de frequência (o módulo), enquanto VAL_QUANT reflecte o número de vezes que o processo ocorre.

O João vai ao ginásio <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="5" VAL_DELTA="A0D7H0M0S0M0"> **cinco vezes por semana** </EM>

O cálculo de VAL_DELTA pode incluir conversão entre unidades. A representação do valor de VAL_DELTA deverá ser o mais próximo possível da expressão presente no texto. Devem ser utilizados apenas valores inteiros e o restante deverá ser convertido para a subunidade imediatamente inferior.

3.5 Atributo UMED

Atributo que representa as unidades de tempo presentes explicitamente na expressão temporal. Se a expressão temporal não envolver unidades de tempo, este atributo é deixado em branco.

Isso aconteceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" VAL_DELTA="A0D1H0M0S0M0" SENTIDO="POSTERIOR" UMED="dia">**no dia seguinte**</EM>

3.6 Atributo FUZZY

O atributo FUZZY está presente com os atributos VAL_NORM e VAL_DELTA, quando um TIMEX é modificado por um elemento lexical que o torna impreciso ou vago.

O João fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+C------D20090423T---------E--LM-" FUZZY="V"> **por volta de 23 de Abril de 2009**</EM>

4 Implementação

Neste capítulo é descrita a implementação do módulo de processamento de ET desenvolvido nesta dissertação e as principais decisões tomadas ao longo da mesma. Este capítulo encontra-se dividido em duas secções: uma secção correspondente ao trabalho realizado no XIP, para a identificação e classificação das ET e uma segunda secção onde é descrito o módulo de pós-processamento desenvolvido e que é responsável pela normalização das ETs anteriormente identificadas.

4.1 Identificação e Classificação de ET

A identificação e classificação de ET já era anteriormente realizada pelo XIP, contudo pretendia-se com este trabalho realizar a identificação e classificação de um número mais abrangente de ET. Por outro lado, uma vez que estes processos eram feitos maioritariamente nas gramáticas locais com recurso a regras específicas, devido à sua especificidade, o número de regras nas gramáticas locais do tempo era significativamente grande. A grande quantidade de regras existentes tornava difícil perceber o encadeamento das regras e dificultava a adição ou alteração das regras já existentes.

Uma das principais dificuldades no reconhecimento das ET é o elevado conjunto de combinações de elementos lexicais que podem ser associados às expressões temporais, a diversidade de valores semânticos que estão presentes numa ET e o facto de algumas ET serem vagas e difíceis de identificar. Assim, ao aumentarmos o número de expressões temporais que se pretendem identificar estaríamos também a aumentar significativamente o número de regras utilizadas para a sua identificação.

Tendo em conta este facto, adoptou-se neste trabalho uma abordagem diferente da anterior. Em vez de se utilizar regras específicas nas gramáticas locais, deixou-se o chunker actuar com regras de tokenização globais e, posteriormente, agrupou-se os diferentes chunks (sintagmas

nucleares) já identificados quando estes constituem uma ET. Pretende-se, assim, minimizar o número de regras utilizadas e a especificidade das mesmas.

A abordagem seguida e as mudança nas directivas do TEMPO criou ainda necessidade de eliminar algumas regras, alterar as regras existentes e reorganizar os ficheiros de TEMPO existentes.

De seguida, apresenta-se as diferentes etapas na identificação e classificação de expressões temporais.

4.1.1 Léxicos

O léxico do TEMPO é compostos por dois ficheiros:

- LEXTime.xip - Neste ficheiro são adicionados os traços que permitem a identificação dos meses do ano, os anos, os dias da semana, as partes do dia, as unidades de tempo, palavras de de algum modo denotam o tempo, etc. Considere-se os seguintes exemplos de entradas lexicais presentes neste ficheiro:

```
janeiro:  noun +=[t-month=1, time=+].
```

Figura 4.1: Exemplo de uma entrada lexical existente no ficheiro LEXTime.xip

Nesta regra, a um nó noun com o lema "`janeiro`", são adicionados os traços `time` e `t-month`; neste caso o valor atribuído a `t-month` (traço utilizado para identificar os diferentes meses do ano) toma o valor 1, dado que se trata do primeiro mês do ano.

No segundo exemplo (ver Figura 4.2), são adicionados ao lema "`antes de ontem`"os traços `time`, `t-val`, `t-tempref` e `timeref`. Uma vez mais, o traço genérico `time` identifica expressões como pretencendo ao léxico do tempo; o atributo `t-tempref` toma o valor `enunc`, correspondente a uma ET referencial cuja interpretação depende do momento de enunciação; finalmente o traço `t-val` toma o valor 2 e `timeref` um valor negativo, o que corresponde, respectivamente à distância temporal entre o tempo do evento indicado pela expressão temporal e o momento de referência, e o sentido dessa distância, neste caso dias, devem ser subtraídos ao momento de referência.

```
"antes de ontem" += [t-val=2, t-tempref=enunc, time=+, timeref=-].
```

Figura 4.2: Exemplo de uma entrada lexical existente no ficheiro LEXTime.xip

- LEXTimeFestive.xip - Efectuou-se um levantamento das expressões que designam festividades, feriados nacionais, feriados religiosos, feriados municipais, quadras festivas, períodos históricos, eras e dinastias. Todas estas expressões foram adicionadas como novas entradas no léxico. Além de este levantamento permitir a identificação/classificação destas expressões como ET, também vai permitir a posterior normalização de algumas destas datas. Tome-se, como exemplo a expressão "dia de Natal":

```
"dia de natal"  += [time=+, t-date=+, t-tempref=enunc, t-one-day=+].
```

Figura 4.3: Exemplo de uma entrada léxical presente no ficheiro LEXTimeFestive.xip

A expressão linguística "dia de Natal" é uma expressão composta, anteriormente identificadas pelo módulo de análise morfológica e à qual vão ser adicionados os traços `time`, `t-date`, e `t-tempref` este último com o valor `enunc`. Estes traços vão permitir determinar que se trata de expressão temporal, do tipo `TEMPO_CALEND` subtipo `DATA`, que corresponde a um tempo de referência de dependente do momento de enunciação e que é uma festividade com a duração de 1 dia e não um período de tempo de outra natureza ou extensão.

4.1.2 Gramáticas Locais

Devido à abordagem seguida, houve necessidade de reestruturar os ficheiros das gramáticas locais. Grande parte das regras anteriormente existentes foram eliminadas e as restantes foram redistribuídas pelas seguintes ficheiros, que são invocados pela seguinte ordem:

`LGTimeNoun.xip >> LGTimeHours.xip >> LGTimeAdv.xip`

O critério utilizado para eliminar as regras anteriormente existentes nas Gramáticas Locais do TEMPO foi o seguinte: Como foi referido anteriormente, a nova estratégia passa por deixar primeiro actuar o chunker e só depois agrupar os chunks para extrair a ET.

- LGTIMENoun.xip - Neste ficheiro são criados nós do tipo `noun` que correspondem a nomes considerados como expressões compostas (séculos/eras/milénios). Por exemplo, a

expressão "II milénio a.C.". A regra vai criar um nó do tipo `noun` com os traços, `time`, `t-date` e `t-tempref` com o valor absoluto, sempre que encontra um número romano ou cardinal, seguido pela palavra milénio, seguido opcionalmente pela abreviatura a.C. a que corresponde o lema "antes de Cristo".

```
2> noun[time=+,t-date=+,t-tempref=absolut] @= num[rom];num[card, frac:~],
                                              ?[lemma:milénio],
                                             (?[lemma:"antes de Cristo"];
                                              ?[lemma:"depois de Cristo"];
                                              ?[lemma:"ano do Senhor"]).
```

Figura 4.4: Exemplo de uma regra presente na LGTIMENoun.xip

Adicionalmente, este ficheiro que contém regras em que são marcados os números que estão associados a unidades de tempo. Posteriormente estes traços vão ser utilizados pelo módulo de normalização. Considere-se o seguinte regra (Figura 4.4) como exemplo:

```
1> ? @= num[t-second=+];?[lemma:um,t-second=+] noun[lemma:segundo,masc].
```

Figura 4.5: Exemplo de uma regra presente na Gramática Local LGTIMENoun.xip

Em que um número seguido por um `noun` com o lema `segundo` é marcado com o traço `t-second` que identifica esse valor como correspondendo à unidade de tempo do segundo, posteriormente, este traço vai ser utilizado no processo de normalização das expressões temporais.

- LGHours.xip - contém as regras que se constroí os nós a que corresponde a unidade de tempo das horas. As horas foram consideradas um caso especial, no critério de eliminação das regras anteriormente descrito. Mantiveram-se as regras que agrupavam as horas nas gramáticas locais, devido ao elevado número de formatos diferentes que as horas podem tomar. Adicionalmente, estas regras permitem marcar as diferentes expressões numéricas integradas nas expressões temporais nas expressões com os traços `t-hour`, `t-minute`, `t-second`, `t-milisecond`, que mais tarde irão ser utilizadas na normalização. Considere-se o exemplo de uma regra que cria um nó `noun` para expressões temporais que denotam horas, e que identifica expressões como "15 minutos antes das 2 horas da madrugada". Repare-se que esta regra, além de criar um nó `noun` com a expressão temporal, são indentifica as expressões numéricas associadas aos valores das unidades minuto e hora presentes na

expressão, às quais são atribuídos respectivamente os traços `t-minute`, e `t-hour`.

```
3> noun[t-date=+,t-hora=+,time=+] @= num[frac:~,ord:~,sem-measother:~,
                                        t-minute=+],
                                   ?[lemma:minuto],
                                   ?[lemma:"antes de"];?[lemma:"depois de"],
                                   ?[surface:a];?[surface:as],
                                    num[frac:~,ord:~,sem-measother:~,
                                        t-hour=+];
                                   ?[lemma:um,num=+,t-hour=+],
                                  (?[lemma:hora];?[lemma:h]),
                                  (?[lemma:"antes do meio-dia"]),
                                  (prep[lemma:de], ?[surface:a],
                                  ?[lemma:manhã];?[lemma:tarde];
                                  ?[lemma:noite];?[lemma:madrugada]).
```

Figura 4.6: Exemplo de uma regra da Gramática Local LGHours.xip

LGTimeAdv.xip - neste ficheiro são criados nós do tipo `adv` referentes a frequências e durações. A regra seguinte Figura 4.7 cria um nó do tipo adv, sempre que encontra duas unidade de tempo ou estações do ano, desde que tenham lemmas iguais.

```
1> adv[advtimefreq=+]@=(?[lemma:de]),
                       #1[t-meas];#1[lemma:momento];#1[t-season],
                       ?[surface:a];?[surface:em],
                       #2[t-meas];#2[lemma:momento];#2[t-season],
                       where(#1[lemma]:#2[lemma]).
```

Figura 4.7: Exemplo de uma regra presente na Gramática Local LGTimeAdv.xip

4.1.3 Chunker

Após serem chamados os ficheiros do Chunker que agrupam os diferentes tokens com recursos a regras gerais, são então chamados os ficheiros específicos para o chunking do TEMPO.

- ChunkerTime1 - são propagados os traços existentes nos tokens para os nós que foram anteriormente criados pelo Chunker. Considere-se o exemplo na Figura 4.8:

```
| NP[t-several-days=+]{art, noun[t-several-days]} | ~
```

Figura 4.8: Exemplo de uma regra existente no ChunkerTime1.xip

A importância da propagação dos nós prende-se não só com a sua normalização mas também com o facto de os traços serem posteriormente utilizados pelo ChunkerTime2 para agrupar nós.

- ChunkerTime2.xip - Neste ficheiro são agrupados os nós previamente constituídos pelo chunker e são formados novos nós, que vão delimitar a expressão temporal de acordo com as directivas adoptadas, utilizando-se para tal os traços existentes nos nós já construídos.

```
1> NP[time=+, t-date=+, t-tempref=absolut] = NP[t-monthday], PP[t-month],
                                              PP[t-year].
```

Figura 4.9: Exemplo de regra presente no ChunkerTime2.xip

Considere-se a regra apresentada na Figura 4.9, que agrupa numa entidade única, expressões do tipo "dia 23 de Maio de 2003". Os nós anteriormente criados pelo chunker como dia do mês, mês e ano são agrupados num nó único nó `NP`. A este nó são atribuídos traços que permitem a classificação desta expressão como sendo do tipo `TEMPO_CALEND` e dp subtipo `DATA`, com um tempo de referência `absoluto`.

Este ficheiro encontra-se organizado por camadas. Nas primeiras, agrupam-se expressões que correspondem a datas ou durações simples. As camadas seguintes vão, por sua vez, agrupar expressões compostas, isto é, são constituídas por combinações dos nós que foram anteriormente construídos e classificados. Por exemplo, expressões temporais do tipo `DATA` subtipo `COMPLEXO`, uma vez que este subtipo particular envolve tanto uma data como uma duração. A regra presente na figura 4.10 constrói um nó `ADVP` para expressões do tipo "de 27 de Maio a 15 dias":

```
23>ADVP[t-complex=+,time=+,t-tempref=enunc]=PP[time,t-date,t-tempref:enunc],
                                            PP[time,t-duration,t-meas].
```

Figura 4.10: Exemplo de uma regra presente no ChunkTime2

Já as expressões do tipo `DATA`, e do subtipo `INTERVALO`, são constituídas por dois limites temporais, ou seja, duas datas (por exemplo, "de 20 de Dezembro a 30 de Janeiro"). Nesta regra (Figura 4.10), constrói-se um nó formado por duas datas.

Note-se que esta nova abordagem para identificar/classificar ET mudou a estrutura dos nós: em que anteriormente se estruturava uma frase, agora, uma expressão temporal pode

```
23> ADVP[t-interval=+, time=+, t-tempref=enunc] = PP[ time, t-date, t-tempref:enunc],
                                                   PP[time, t-date, t-tempref:enunc].
```

Figura 4.11: Exemplo de uma regra presente no ChunkTime2

resultar do agrupamento de diversos nós, ao contrário do que acontecia anteriormente, quando as expressões eram identificadas nas gramáticas locais (Figura 4.12). A nova abordagem leva a uma maior simplicidade e modularidade na formulação das regras no Chunker, tornando mais fácil a manutenção e revisão da gramática.

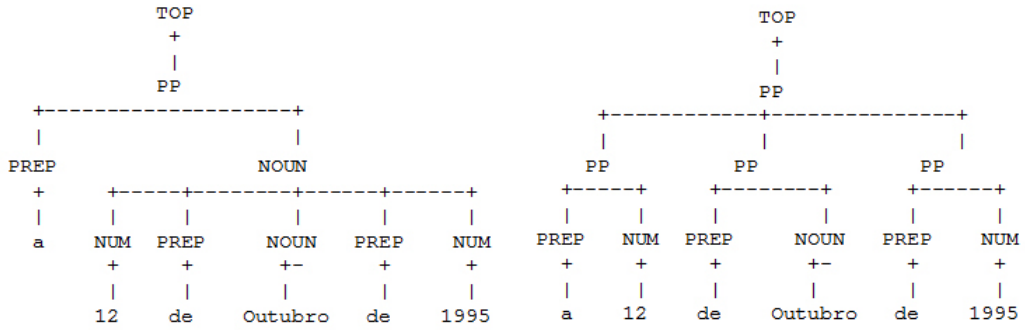


Figura 4.12: Exemplos da estrutura de nós `TEMPO`. Estrutura dos nós segundo a abordagem anterior - Estrutura dos nós segundo a nova abordagem.

4.1.4 Dependências

- Entit_dependencyTime.xip - é realizada a classificação e identificação final consoante os traços existentes nos nós anteriormente criados, são criadas as diferentes entidades de TEMPO. Considere-se o seguinte exemplo da classificação de uma entidade temporal com o subtipo `COMPLEXO`(Figura 4.13):

```
| ADVP#1[time,t-complex, t-tempref:enunc] |
if ( ~$1(#1) )
NE[tempo=+,t-complex=+, t-ref-enunc=+, t-ref-simult=+](#1)
```

Figura 4.13: Exemplo de uma regra presente no Entit_dependencyTime

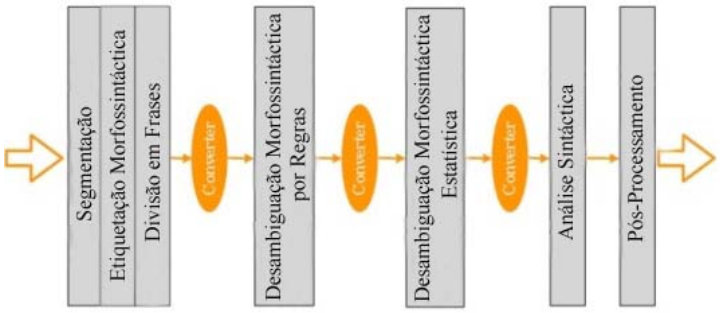


Figura 4.14: Arquitectura da Cadeia de Processamento do L2F

4.2 Normalização de ET

Anteriormente a normalização de expressões temporais era realizada dentro do XIP com recurso ao Python, contudo a normalização não era utilizada por nenhuma regra do XIP. Ao se deslocar a normalização para um módulo exterior ao XIP, pretendeu-se eliminar a dependência ao Python, evitando assim os problemas de instalação ocorridos anteriormente sempre que surgiam novas versões do Python.

Na Figura 4.14, encontra-se representada a cadeia de processamento do L^2F e o módulo de pós-processamento que terá como entrada a saída em XML do XIP.

4.2.1 Estrutura do Módulo de Processamento

A normalização de ETs é realizada pelo módulo de pós-processamento implementado em Java e que tem como entrada a saída XML do XIP.

4.2.1.1 Input do Módulo de Normalização de ET

O input do módulo de normalização de ETs, é um ficheiro de saída do XIP, que possui os seguintes elementos na sua estrutura:

- `XIPRESULT` - que contêm uma ou mais `LUNITS`;
- `LUNIT` - corresponde a uma unidade linguística, em que cada elemento deste tipo vai corresponder a uma frase, pode conter um conjunto de `NODES` e de `DEPENDENCY`;
- `NODE` - corresponde ao resultado da análise morfossintática. Pode conter outros `NODES` ou `TOKENS`;

- `TOKEN` - corresponde ao resultado da tokenização;
- `DEPENDENCY` - corresponde a relações sintáticas entre os nós.

Note-se que cada um destes elementos possui ainda vários atributos.

4.2.1.2 Normalizador

O normalizador de ETs utiliza a XIP APPI desenvolvida por Nuno Nobre (Nobre 2011). A XIP APPI manipula o ficheiro de entrada e constrói um objecto do tipo `XipDocument`. Este objecto contém uma árvore de chunks constituída por objectos do tipo `XIPNodes`.

Posteriormente o Normalizador de ETs utiliza o XipDocument. Sempre que é encontrado um nó que foi classificado como uma entidade temporal, o normalizador cria objectos que representam os diferentes subtipos de entidades temporais: datas, intervalos, expressões de tempo complexas, expressões de duração simples e do tipo intervalo e, finalmente, expressões de frequência.

Cada uma das classes que representa as entidades temporais, possui métodos que identificam as unidades a normalizar, converte os valores e constroem os valores a devolver. Salienta-se, mais uma vez, a importância dos traços anteriormente atribuídos, que permitem identificar as componentes da expressão temporal que são normalizáveis. Também na normalização se utilizou o facto de algumas entidades de tempo serem expressões compostas, por exemplo datas do tipo `COMPLEXO` que são compostas por uma data e uma duração.

4.2.2 Output do Módulo de Processamento

A saída deste módulo corresponde a um ficheiro com os mesmos elementos que o ficheiro de entrada. No entanto, cada elemento `LUNIT` pode conter um ou mais elementos do tipo `TIME_NORMALIZATION` que representa a normalização do TEMPO.

5 Avaliação

Neste capítulo é apresentada a metodologia utilizada na avaliação do sistema de processamento de expressões temporais desenvolvido no âmbito desta dissertação, assim como os resultados obtidos na avaliação.

Durante a realização deste trabalho utilizou-se um corpus de desenvolvimento para testar o progresso do trabalho. Este corpus é composto por cerca de 1800 frases que contêm expressões temporais. Realça-se que este corpus de desenvolvimento permitiu a identificação de alguns problemas durante a fase de desenvolvimento.

Para avaliação deste trabalho, utilizou-se um corpus diferente, a Colecção Dourada, que foi o corpus de avaliação utilizado no Segundo HAREM. Este corpus foi anotado por uma linguista do L2F, de acordo com as directivas desenvolvidas nesta dissertação.

5.1 Caracterização do Corpus de Avaliação

Na anotação manual do corpus de avaliação, feita de acordo com as novas directivas desenvolvidas no âmbito desta dissertação, foram dentificadas 1250 entidades do tipo `TEMPO`. Na tabela seguinte encontra-se a distribuição destas entidades pelos diferentes tipos e subtipos.

TIPO	**SUBTIPO**	
TEMPO_CALEND	DATA	1016
	INTERVALO	73
	COMPLEXO	1
	RESIDUAL	0
DURACAO	SIMPLES	65
	INTERVALO	6
FREQUENCIA		72
GENERICO		17

Tabela 5.1: Distribuição das Entidades `TEMPO` no corpus de Avaliação

Note-se que de acordo com as directivas seguidas no Segundo HAREM na CD foram identificadas 1195 entidades de tempo, sendo 974 do tipo TEMPO_CALEND, 71 do tipo FREQUENCIA, 60 do tipo DURACAO e 90 do tipo GENERICO. Podemos verificar que com as directivas seguidas nesta dissertação o número de ET identificadas aumentou em 55 entidades. As principais diferenças prendem-se com o número de entidades identificadas do tipo TEMPO_CALEND e GENERICO. O número de entidades do tipo GENERICO diminuiu significativamente, tendo apenas sido identificadas nas directivas actuais 17 entidades. Enquanto que o número de entidades TEMPO_CALEND aumentou em 116 entidades.

5.2 Métodos de Avaliação e Métricas Utilizadas

Com o objectivo de avaliar diferentes aspectos na identificação e classificação de ET e por forma a aproximar esta avaliação da realizada no Segundo HAREM, foram realizados três tipos de avaliação distintos:

- **Identificação da Categoria TEMPO** onde se pretende avaliar apenas a delimitação e identificação das ET.
- **TEMPO Clássico**, onde apenas são considerados os atributos CATEG, TIPO e SUBTIPO.
- **TEMPO Clássico com o Atributo TEMP_REF**, onde são considerados os atributos CATEG, TIPO, SUBTIPO e TEMP_REF. O atributo SENTIDO não foi contemplado nesta avaliação uma vez que a anotação manual do corpus não continha a anotação deste atributo.

Adicionalmente, foi também avaliado o módulo de pós-processamento reponsável pela normalização das ETs. Com o objectivo de avaliar a normalização das expressões temporais, foram considerados os atributos que estão associados à normalização (VAL_NORM, VAL_DELTA, UMED e FUZZY). Uma vez que a normalização é realizada num módulo de pós-processamento exterior ao XIP, a avaliação dos atributos referentes à normalização foi efectuada separadamente dos restantes processos.

5.2.1 Métricas Utilizadas

Para os diferentes modos de avaliação da identificação e classificação, a saída do sistema, foi comparada com a anotação manual do corpus. As ET identificadas podem ser consideradas `CORRECTA`, `ESPURIA` ou `EM FALTA`. São consideradas `CORRECTAS` as ET identificadas pelo sistema que são iguais às ET anotadas manualmente. Quando é identificada pelo sistema uma ET que não se encontra na anotação manual do corpus, essa ET é considerada `ESPURIA`. São consideradas ET `EM FALTA` se o sistema não identifica a ET ou se apenas identifica parcialmente os seus atributos.

Para as diferentes formas de avaliação serão calculadas as seguintes métricas:

$$Precisão = \frac{Expressões\_Temporais\_Correctas}{Expressões\_Temporais\_Identificadas} \tag{5.1}$$

$$Abrangência = \frac{Expressões\_Temporais\_Correctas}{Total\_de\_Expressões\_Temporais} \tag{5.2}$$

$$Medida - F = \frac{2 * Precisão * Abrangência}{Precisão + Abrangência} \tag{5.3}$$

As métricas anteriores também foram utilizadas no Segundo HAREM para avaliar os diversos participantes. Contudo, não se pode realizar uma comparação directa dos resultados deste trabalho com os resultados obtidos pelos diversos participantes no Segundo HAREM, uma vez que as directivas sofreram algumas modificações e não são, por isso, directamente comparáveis.

Para a avaliação do módulo de pós-processamento responsável pela normalização de ETs, utilizaram-se apenas as expressões temporais presentes na CD que foram classificadas correctamente pelo sistema, uma vez que essa classificação vai ser o ponto de partida para a normalização de expressões temporais. Para avaliar o módulo de pós-processamento de utilizou-se as seguintes métricas:

$$Precisão = \frac{Expressões\_Temporais\_Correctamente\_Normalizadas}{Expressões\_Temporais\_Normalizadas\_pelo\_Sistema} \quad (5.4)$$

Esta métrica mede a proporção de respostas correctas em todas as respostas fornecidas pelo módulo de pós-processamento. Não foram consideradas ET correctamente normalizadas quaisquer expressões para as quais o sistema apresentou uma saída apenas parcialmente correcta.

5.3 Resultados

Nesta subsecção apresentam-se os resultados obtidos nos diferentes modos de avaliação.

5.3.1 Identificação de Expressões Temporais

Na Tabela 5.2 são apresentados os resultados na identificação das Expressões Temporais.

Precisão	**Abrangência**	**Medida-F**
0,6667	0,7424	0,7024

Tabela 5.2: Resultados obtidos para a Identificação da Categoria `TEMPO`

Neste modo de avaliação, obteve-se uma precisão de 0,6667 e uma abrangência de 0,7424. Analisando os resultados obtidos, identificaram-se os principais problemas na identificação/delimitação de ETs:

- Verificou-se a existência de expressões temporais que ainda não são identificadas pelo sistema como o caso de "Ano Mundial da Astronomia" "era Elizabetana". No âmbito deste trabalho, realizou-se um levantamento de expressões que representam períodos e épocas históricas, contudo, este deve ser ainda enriquecido e deverão ser adicionadas novas regras às gramáticas locais, que permitam a identificação de expressões que não estejam ainda presentes no léxico.
- Expressões que foram erradamente consideradas como ET pelo sistema, como por exemplo "de altura" "na altura". Repare-se que algumas destas expressões podem fazer parte de expressões temporais (por exemplo, "na altura do Natal"), contudo estas

expressões sozinhas ou a expressão"de altura"não devem ser consideradas expressões temporais.

- Situações em que o sistema identificou como expressões temporais, títulos de obras literárias, (por exemplo,"Se Numa Noite de Inverno um Viajante"). Repare-se que nesta situação, "numa noite de Inverno"não representa uma ET mas o título da obra é formado exactamente sobre a matriz da expressão temporal.
- Verificou-se a delimitação incorrecta de algumas expressões temporais. Por exemplo, a expressão "desde, pelo menos, o século 19"em que o sistema apenas identificou "século 19"não tendo sido capaz de tratae a inserção do abverbial "pelo menos"; ou a expressão "nos idos dos anos 50"em que apenas se identificou "dos anos 50"como ET trata-se de uma expressão em que os "idos"não é usado no seu sentido próprio mas figurativamente, já que, em rigor e literalmente, este nome se refer à forma latina de marcação do texto e apenas admite como complemento nomes de meses. A delimitação incorrecta dos exemplos anteriores resulta de não se ter contemplado estas combinações de elementos temporais aquando do agrupamento dos diferentes elementos que constituem a ET.

5.3.2 TEMPO CLÁSSICO

Na Tabela 5.3 encontram-se representados os diferentes resultados obtidos para a classificação dos tipos e subtipos da categoria do `TEMPO`.

Precisão	**Abrangência**	**Medida-F**
0,6625	0,7312	0,6951

Tabela 5.3: Resultados obtidos para a classificação dos `TIPO` e `SUBTIPO`

Os resultados obtidos para a classificação dos tipos e subtipos são ligeiramente inferiores aos resultados obtidos apenas na classificação da entidade temporal. Podemos, assim, concluir que os principais problemas existentes e que têm um impacto mais significativo nos resultados, se prendem com a identificação e delimitação das ET, os quais foram anteriormente descritos, e não tanto com a classificação das entidades nos seus diferentes tipos e subtipos.

De forma a perceber como o sistema se comporta na classificação dos diferentes `TIPOS` da

categoria TEMPO foram calculadas as mesmas métricas mas para os diferentes TIPOS (ver Tabela 5.4). De seguida, é realizada uma descrição dos principais problemas identificados para cada subtipo da categoria TEMPO.

TIPO DE ET	Precisão	Abrangência	Medida-F
TEMPO_CALEND	0,7284	0,7371	0,7327
DURACAO	0,4915	0,4573	0,47373
FREQUENCIA	0,6521	0,6164	0,6338
GENERICO	0,0632	0,5556	0,1136

Tabela 5.4: Resultados obtidos para a classificação dos diferentes tipos

5.3.2.1 TIPO TEMPO_CALEND

O tipo TEMPO_CALEND é o tipo que obteve melhores resultados na avaliação diferenciada dos diferentes tipos para a CATEGORIA TEMPO, com uma precisão de 0,7284 e uma abrangência de 0,7371. É de salientar que este tipo é igualmente o mais comum e o que se encontra melhor representado na CD. Identificaram-se os principais problemas na classificação deste tipo:

- A não identificação de alguns períodos de tempo, (por exemplo: "durante os desolados anos Regan", "era Elizabetana"tal como anteriormente foi referido. Verificou que apesar do léxico apresentar uma longa listagem de eras e periodos, não é suficiente, existe a necessidade de complementar o léxico com regras nas gramáticas locais.
- A não identificação de alguns número que representam anos, principalmente nas situações em que datas isoladas aparecem entre parênteses. A identificação de anos é especialmente complexa, porque 1995 tanto pode representar um ano como pode representar uma quantidade, por isso na identificação dos anos é utilizado o contexto, o que torna as datas, anos isoladas, particularmente difíceis de identificar.
- A identificação de intervalos do tempo com os seguintes formatos : "2003 - 2006"e "5/6 de Outubro". O sistema apenas identifica o subtipo intervalo com o seguinte formato "de 2003 a 2006 "e "entre 2003 e 2006", não identificando os anteriores formatos. É de salientar a ocorrência relativamente frequente deste tipo de intervalo.

– Outra situação de não identificação aconteceu com expressões como “deste ano”, “deste dia”, “deste Inverno”, que deviam ter sido assinaladas como ET.

5.3.2.2 TIPO DURACAO

Quanto ao tipo `DURACAO`, verifica-se que este apresenta resultados de precisão e abrangência inferiores aos resultados gerais. Note-se que para este subtipo apenas foram identificadas manualmente 71 entidades na CD, o que leva a que o impacto de uma ET classificada erradamente seja bastante maior do que no caso das datas. Os principais problemas que foram identificados na classificação de expressões temporais do `TIPO DURACAO` foram os seguintes:

– A existência de algumas durações que foram incorrectamente classificadas como `TEMPO_CALEND`.

– A incorrecta delimitação de expressões como “por mais de um minuto”e “por mais de um ano”, em que apenas se identificou e classificou como sendo `DURACAO` a expressão “um minuto”e “um ano”.

– A incorrecta delimitação de expressões como “num prazo de 20 minutos”, já que durante a construção de regras não tenha sido este caso considerado.

5.3.2.3 TIPO FREQUENCIA

Analisando os resultados obtidos, verifica-se que este TIPO apresenta resultados de precisão e abrangência semelhantes aos resultados gerais. Foram identificadas as seguintes situações em que a classificação deste tipo não é feita correctamente:

– A incorrecta identificação de expressões como frequência como “imediatamente”. Efectivamente, este advérbio é uma ET, contudo , deveria ter sido classificado como data referencial (textual).

– A não identificação de expressões “por ano” “por mês” “de tanto em tanto tempo”, “quotidianamente”

5.3.2.4 TIPO GENERICO

O tipo GENERICO, é o tipo dentro da categoria do TEMPO que apresenta os piores valores. Este tipo de entidade é sem dúvida o mais difícil de identificar, uma vez que não representa precisamente uma ET mas é, no entanto, composto por elementos léxicais com valor semântico temporal. A estratégia seguida na identificação e classificação deste tipo de ET foi a seguinte: no cálculo das dependências do tempo (ficheiro dependencyTime.xip) o sistema faz a classificação final de todas as datas, durações e frequências. Posteriormente, as expressões que contenham elementos temporais e a que não tenha sido atribuída nenhuma classificação são identificadas como sendo do tipo GENERICO. Isto leva a algumas expressões, que deveriam ter sido classificadas como TEMPO_CALEND, FREQUENCIA e DURACAO, acabem sendo classificadas indevidamente como pretencendo ao tipo GENERICO.

Note-se, porém, que no corpus há apenas 18 ET deste tipo, o que faz com que, quantitivamente, estes resultados tenham impacto reduzido nos resultados globais.

5.3.3 TEMPO CLÁSSICO com o atributo TEMPO_REF

Entre as entidades identificadas como TEMPO_CALEND, 581 apresentam um tempo de referência ABSOLUTO o que corresponde 54% das entidades anotadas com o atributo TEMPO_REF, 368 exibem um tempo de referência com valor ENUNCIACAO (34%) e 134 têm um tempo de referência com valor TEXTUAL (12%).

Comparado a distribuição dos valores do atributo TEMPO_REF com a distribuição deste atributo com as directivas utilizadas no Segundo HAREM verifica-se que a distribuição é semelhante, sendo o valor com maior representação o valor ABSOLUTO, seguido pelo valor ENUNCIACAO e finalmente o valor TEXTUAL.

Precisão	**Abrangência**	**Medida-F**
0,9662	0,7233	0,8272

Tabela 5.5: Resultados obtidos para a classificação dos TIPO, SUBTIPO e atributo TEMPO_REF

Na Tabela 5.5 são apresentados os resultados da classificação do tipo e subtipo e o atributo TEMPO_REF. Nesta forma de avaliação obteve-se uma precisão de 0,9662 e uma abrangência de 0,7233.

5.3.4 Avaliação do Módulo de Pós-Processamento

Tal como foi anteriormente referido, para a avaliação do módulo de pós-processamento, utilizaram-se as expressões temporais presentes na CD que foram classificadas correctamente pelo sistema, uma vez que essa classificação vai ser o ponto de partida para a normalização de expressões temporais. Na Tabela 5.6 encontram-se os resultados obtidos na avaliação do módulo de normalização para os tipos `TEMPO_CALEND`, `DURACAO` e `FREQUENCIA`.

TIPO	**Precisão**
TEMPO_CALEND	0,9012
DURACAO	0,7778
FREQUENCIA	0,8048

Tabela 5.6: Resultados obtidos na Avaliação do Módulo responsável pela Normalização

Para o tipo `DURACAO`, foram considerados os atributos `VAL_DELTA`, `UMED` e o atributo `FUZZY`. Para o subtipo `FREQUENCIA`, foram considerados os atributos `VAL_QUANT`, `VAL_DELTA` `UMED` e o atributo `FUZZY`.

Os resultados obtidos demonstram que a normalização das expressões temporais é feita em grande partes dos casos correctamente. Ainda assim, foram detectados alguns problemas que, futuramente, deverão ser solucionados: trata-se da incorrecta conversão de alguns números expressos por extenso e a inadequada normalização de durações do sutipo `INTERVALO`, por exemplo "entre 3 e 4 horas". Tal como foi anteriormente referido, as durações do tipo `INTERVALO` são tratadas pelo normalizador, como sendo 2 durações simples, "3"e "4 horas". Para que a normalização seja efectuada correctamente, é necessário que seja anteriormente atribuída ao número "3"o traço `t-hour`, para que este possa ser normalizado como sendo horas. .

6 Conclusões e Trabalho Futuro

Ao longo deste trabalho descreveu-se o trabalho desenvolvido no âmbito da identificação, classificação e normalização de expressões temporais, trabalho este integrado na cadeia de processamento do L2F - INESC ID Lisboa.

A identificação e a classificação de expressões temporais não é uma tarefa simples, como anteriormente já foi referido. A grande dificuldade nesta tarefa prende-se como o elevado conjunto de combinações de elementos lexicais que podem ser associados às expressões temporais e com a diversidade de valores semânticos que estão presentes numa ET. Tendo em conta este aspecto e os recentes desenvolvimentos na cadeia de processamento, modificou-se a abordagem seguida (descrita em detalhe no Capítulo 4) na identificação e delimitação de expressões temporais dando maior importância ao módulo de chunking. Tal corresponde a tornar consequentemente a observação que a maioria das ET obedece de um modo geral às regras de boa formação sintáctica de constituintes da Língua. Esta abordagem mostrou ser uma abordagem válida, levando à diminuição de regras nas gramáticas locais, tendo, no entanto, levado a um aumento do número de regras no chunkerTime. Contudo estas regras foram desenvolvidas de forma modular, o que facilita a tarefa de manutenção e a actualização da gramática, nomeadamente, a adição e alteração de novas regras no futuro.

Os resultados obtidos na avaliação deste trabalho são positivos e encorajadores. No entanto, demonstram que ainda podem ser feitas melhorias quer no âmbito da identificação, quer no âmbito da classificação das ETs. Foram identificadas as situações problemáticas e a correcção das mesmas deverá levar a uma melhoria significativa na resposta do sistema.

Agora que é realizada a identificação, a classificação e a normalização das ETs, um trabalho futuro relevante seria permitir que os diversos eventos mencionados num texto podessem ser ordenados numa sequência cronológica. Tal não é uma tarefa trivial, já que é necessário, em primeiro lugar, associar os eventos (estados, acções, processos) às ET identificdas e anotadas, bem como resolver as referências temporais.

Apesar de as directivas serem abrangentes na classificação e permitirem a normalização de grande parte das expressões temporais, foram identificadas algumas limitações às directivas aqui seguidas. De seguida, são feitas algumas sugestões de trabalho futuro, tendo como finalidade ampliar o conjunto de expressões temporais normalizáveis.

As directivas adoptadas apenas contemplam a normalização de frequências simples. Em trabalhos futuros sugere-se que este tipo seja desdobrado em dois subtipos: um subtipo `SIMPLES`, que abrange expressões com "2 vezes por semana"já descrito e normalizado no quadro das directivas actuais e um novo subtipo `COMPLEXO`, correspondendo a expressões como "8 horas por dia", para cuja normalização é preciso fazer intervir também um conceito de `DURAÇÃO`.

Nas directivas actualmente adoptadas, os dias da semana e partes do dia não são normalizados, pelo que se propõe uma alteração ao formato do atributo `VAL_NORM` (Figura 6.1), de forma a poder normalizar estas expressões. O valor de <DiaDaSemana> seria um digito de 1 a 7 que representaria o dia da semana e <ParteDia> uma expressão que representaria as partes do dia (por exemplo, `AM` corresponderia amanhecer, `TA` corresponderia a tarde, etc.)

```
VAL_NORM="<Era>M<Milenio><Seculo><Decada>D<Ano><Mes><Dia>
          T<Hora><Minuto><Segundo><Milissegundo>E<Estacao>
          DS<DiaDaSemana>PD<ParteDia> LM<limite_aberto>
```

Figura 6.1: Proposta para o atributo VAL_NORM

Outro aspecto importante a considerar prende-se com o cálculo de datas referênciais. Actualmente as datas referênciais são identificadas pelo atributo `TEMPO_REF` com os valores `TEXTUAL` e `ENUNCIACAO`. Estas poderiam ser calculadas mediante o conhecimento do momento de enunciação ou a identificação no texto da data de referência.

Referências

Amaral, C., H. Figueira, A. Mendes, P. Mendes, C. Pinto, & T. Veiga (2008). *Adaptação do Sistema de Reconhecimento de Entidades Mencionadas da Priberam ao HAREM*, Chapter 9 in Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM, pp. 171–179.

Baptista, J., N. Mamede, C. Hagège, & A. Maurício (2009). Time expressions in portuguese: Guidelines for identification, classification and normalization. Technical report, L^2F – Laboratório de Sistemas de Língua Falada, INESC-ID Lisboa.

Bick, E. (2000). *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Universidade de Arhus.

Bruckschen, M., J. Guilherme Camargo de Souza, R. Vieira, & S. Rigo (2008). *Sistema SeRELeP para o Reconhecimento de Relações entre Entidades Mencionadas*, Chapter 14 in Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM, pp. 247–260.

Cardoso, N. (2008). *REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto*, Chapter 11 in Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM, pp. 195–211.

Cardoso, N. & D. Santos (2007). Directivas para a identificação e classificação semântica na colecção dourada do HAREM. pp. 211–238.

Craveiro, O., J. Macedo, & H. Madeira (2008). *PorTexTO: Sistema de Anotação / Extracção de Expressões Temporais*, Chapter 8 in Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM, pp. 159–170.

Diniz, C. (2010). Um conversor baseado em regras de transformacão declarativas. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal.

Ferreira, L., A. Teixeira, & J. a. P. d. S. Cunha (2008). *REMMA - Reconhecimento de Entidades Mencionadas do MedAlert*, Chapter 12 in in Desafios na Avaliação

Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM, pp. 213–229.

Ferrucci, D. & A. Lally (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering 10*(3-4), 327–348.

Hagège, C., J. Baptista, & N. Mamede (2008). *Proposta de Anotação e Normalização de Expressões Temporais da Categoria TEMPO para o HAREM II*, Chapter Apêndice B in Desafios na Avaliacao Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM, pp. 289–308.

Hagège, C. & X. Tannier (2007). XRCE-T: XIP temporal module for TempEval campaign. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 492–495. ACL.

Jurafsky, D. & J. H. Martin (2000). *Speech and Language Processing*. Prentice-Hall.

Loureiro, J. a. (2007). Reconhecimento de entidades mencionadas e normalização de expressões temporais. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal.

Mamede, N. (2011). A Cadeia de Processamento XIP. L^2F – Laboratório de Sistemas de Língua Falada.

Martins, B. (2008). *O Sistema CaGE e a Participacão no Segundo HAREM*, Chapter 7 in Desafios na Avaliacao Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM, pp. 149–158.

Martins, B., H. Manguinhas, & J. L. Borbinha (2008). Extracting and exploring the geo-temporal semantics of textual resources. In *Proceedings of the 2th IEEE International Conference on Semantic Computing (ICSC 2008), August 4-7, 2008, Santa Clara, California, USA*, pp. 1–9. IEEE Computer Society.

Medeiros, J. C. (1995). Processamento Morfológico e Correcção Ortográfica do Português. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal.

Nobre, N. (2011). Resolução de Expressões Anafóricas. Master's thesis, Instituo Superior Técnico, – Universidade Técnica de Lisboa, Portugal.

Oliveira, D. (2010). Extraction and classification of named entities. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal.

Pardal, J. P. (2007). Manual do Utilizador do RuDriCo. L^2F – Laboratório de Sistemas de Língua Falada.

Ribeiro, R., N. J. Mamede, & I. Trancoso (2003). Using Morphossyntactic Information in TTS Systems: comparing strategies for European Portuguese. In *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*, Volume 2721 of *Lecture Notes in Computer Science*. Springer.

Romão, L. (2007). Reconhecimento de entidades mencionadas em língua portuguesa. Master's thesis, Instituo Superior Técnico – Universidade Técnica de Lisboa, Portugal.

Santos, D. (2010). Extraccão de relacões entre entidades mencionadas. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal.

Schmid, H. (1995). TreeTagger, a language independent part-of-speech tagger. Technical report, Institut fur Maschinelle Sprachverarbeitung, Universidade de Estugarda.