UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

Prospecção de Conceitos Geográficos na Web

David Soares Batista

MESTRADO EM ENGENHARIA INFORMÁTICA
Especialização em Arquitectura, Sistemas e Redes de Computadores

2009

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

Prospecção de Conceitos Geográficos na Web

David Soares Batista

DISSERTAÇÃO

Projecto orientado pelo Prof. Doutor Mário J. Gaspar da Silva

MESTRADO EM ENGENHARIA INFORMÁTICA
Especialização em Arquitectura, Sistemas e Redes de Computadores

2009

Agradecimentos

Durante a realização e a escrita deste trabalho foram várias as pessoas com quem fui interagindo e que, directa ou indirectamente, contribuiram para o trabalho aqui apresentado. A elas gostava de expressar o meu agradecimento.

Agradeço ao meu orientador, Prof. Mário J. Silva pelo empenho, profissionalismo e dedicação. As revisões constantes do trabalho foram uma mais valia imprescindível. Muito obrigado.

Um agradecimento muito especial ao Francisco J. Lopez-Pellicer por todo o apoio, entusiasmo, sugestões e discussões. Ha sido una motivación constante trabajar con usted! Gracias por todo!

Um agradecimento ao Nuno Cardoso e ao Bruno Martins, embora distantes, as sugestões dadas enriqueceram o meu trabalho.

A todos os meus amigos no LaSIGE, com quem é sempre possível contar.

Aos meus Pais e a todos os meus Amigos

Resumo

Esta dissertação apresenta um estudo feito sobre extracção de informação de documentos, para geração de resumos geográficos. É estudado um método de aprendizagem supervisionada, com base em *Conditional Random Fields* para extracção de entidades em sequências de texto. O método estudado é integrado num sistema desenvolvido no âmbito desta dissertação, o HENDRIX, de forma a poder efectuar a extracção de entidades geográficas para textos em português e o seu tratamento. O tratamento das entidades geográficas extraídas é feito recorrendo a ontologias geográficas. O sistema desenvolvido foi depois usado para fazer a extracção de entidades geográficas de uma colecção de documentos, que representa uma recolha da web portuguesa, sob um *cluster* de computadores.

São apresentados os resultados do desempenho do modelo gerado para extracção de informação geográfica e a análise das entidades geográficas extraídas da recolha da web portuguesa. A partir dos resultados observou-se que o corpus usado para treinar o modelo não é suficientemente expressivo para treinar um modelo de extracção de informação geográfica.

Palavras-chave: extracção de informação geográfica, *conditional random fields*, ontologias geográficas, web semântica

Abstract

This dissertation presents a research done on information extraction for the generation of geographic summaries. The method studied is based on Conditional Random Fields, a supervised learning method for labeling or parsing of sequential data, such as natural language text. This method is then integrated in a system developed during the course of this dissertation. The developed system, HENDRIX, performs geographic entities extraction for documents written in Portuguese. HENDRIX also generates a geographic summary based on the extracted entities and their relations on geographic ontologies. The developed system was then applied to a crawl of the Portuguese Web, using a cluster of computers.

This dissertation presents the results of the performance of the generated model for geographic information extraction as well as an analysis of the extracted entities from the crawl. The results show that the corpus on which the generated model was based is not rich enough to generate a good model for geographic information extraction.

Keywords: geographic information extraction, conditional random fields, geographic ontologies, semantic web

Conteúdo

Lista de Figuras **xv**

Lista de Tabelas **xviii**

1 Introdução **1**
- 1.1 Objectivos . . . 3
- 1.2 Contribuições . . . 3
- 1.3 Metodologia . . . 5
- 1.4 Estrutura do documento . . . 5

2 Trabalho relacionado **7**
- 2.1 Terminologia . . . 7
- 2.2 Text Mining . . . 8
  - 2.2.1 Pré-processamento . . . 8
  - 2.2.2 Categorização e Extracção de Informação . . . 9
  - 2.2.3 Extracção de Informação Geográfica . . . 10
  - 2.2.4 Desambiguação . . . 10
  - 2.2.5 Referências implícitas . . . 12
  - 2.2.6 Conditional Random Fields . . . 12
  - 2.2.7 Medidas de Avaliação . . . 14
  - 2.2.8 Software Analisado . . . 15
- 2.3 HAREM - Avaliação de Reconhecimento de Entidades Mencionadas . . . 19
  - 2.3.1 Colecção Dourada . . . 20
- 2.4 Ontologias Geográficas . . . 22
  - 2.4.1 Geographic Knowledge Base . . . 23
  - 2.4.2 Geo-Net-PT . . . 25
  - 2.4.3 WGO - World Geographic Ontology . . . 25
  - 2.4.4 Wiki WGO 2009 . . . 29
- 2.5 Similaridade Semântica . . . 29
  - 2.5.1 Information Content . . . 30
  - 2.5.2 Medidas de Similaridade Semântica . . . 31

2.6 Sumário 32

3 HENDRIX **33**
3.1 Arquitectura 33
3.2 Geração do modelo CRF 34
3.3 PAREDES 35
3.3.1 Processo de Emparelhamento 36
3.3.2 Processo de Desambiguação 39
3.3.3 Geração de Resumos Geográficos 42
3.4 Processamento de Colecções de Documentos 43
3.4.1 PAGE 43
3.5 Sumário 46

4 Resultados **47**
4.1 Treino do modelo de Reconhecimento de Entidades Geográficas 47
4.2 GikiCLEF 55
4.3 Anotação da WPT05 56
4.3.1 Identificação Linguística 57
4.3.2 Marcação de Entidades Geográficas Mencionadas 60
4.4 Avaliação e Âmbitos Geográficos 63
4.5 Conclusão 67

5 Conclusão e Trabalho Futuro **69**
5.1 Experiências com o modelo CRF 70
5.2 Inferência de Âmbitos Geográficos 70
5.3 Conclusões 71
5.4 Trabalho futuro 72

Abreviaturas **75**

Bibliografia **81**

Índice **82**

Lista de Figuras

1.1 Processo de geração de resumos geográficos 4

2.1 Treino de Conditional Random Fields 13
2.2 Classificação usando um Conditional Random Field 14
2.3 Exemplo de um texto de entrada para o Minorthird 16
2.4 Meta-Modelo do GKB 2.0 23
2.5 Atributos para *Features* e *Types* 24
2.6 Relações entre domínios 24
2.7 Relações entre tipos de conceitos para os dados administrativos 28
2.8 Relações entre tipos de conceitos para os dados físicos 28

3.1 Arquitectura geral do sistema HENDRIX 34
3.2 Arquitectura do módulo PAGE 35
3.3 Expressão regular utilizada para detectar tipos de conceitos 37
3.4 Exemplo de um RDF que descreve um documento 44
3.5 Fluxo de processamento de dados na plataforma HADOOP 45
3.6 Exemplo da saída do processamento de um RDF pelo PAGE 45

4.1 Ocorrências de EM geográficas na CD do HAREM I 48
4.2 Ocorrências de EM geográficas na CD do Mini-HAREM 49
4.3 Ocorrências de EM geográficas na CD do HAREM II 50
4.4 Exemplo do sumário gerado pelo HENDRIX para o GikiCLEF 2009 . . . 56
4.5 Classificação Linguística com base em n-gramas 58
4.6 Distâncias entre dois perfis de n-gramas 59
4.7 Línguas mais frequentes na WPT-05 59
4.8 Expressão regular utilizada para detectar datas 63

Lista de Tabelas

1.1 Entidades extraídas com correspondências nas ontologias 5

2.1 Apelidos correspondentes a nomes de locais 12
2.2 Algumas propriedades do *software* analisado 15
2.3 Outros pacotes de *software* com suporte para CRF 17
2.4 Exemplos de *features* geradas 18
2.5 Categorias e tipos definidos no segundo HAREM 21
2.6 Distribuição de termos segundo a variante de português. 22
2.7 Caracterização Estatística dos Dados Administrativos 26
2.8 Caracterização Estatística dos Dados Físicos 27
2.9 Caracterização Estatística dos dados administrativos na WGO 29
2.10 Caracterização Estatística dos dados físicos na WGO 30

3.1 Caracterização das CD para a categoria LOCAL 35
3.2 Exemplo de representações de nomes alternativos nas ontologias do sistema GKB. 36
3.3 Exemplos de abreviaturas expandidas. 37
3.4 Separação entre o tipo de conceito e o seu nome 38

4.1 Entidades geográficas mais frequentes para a CD do HAREM I 49
4.2 Entidades Geográficas mais frequentes para a CD do Mini-HAREM . . . 50
4.3 Entidades geográficas mais frequentes para a CD do HAREM II 51
4.4 Identificação de EM da categoria LOCAL no HAREM II 52
4.5 Distribuição das funções de característica pelas etiquetas de classificação 52
4.6 Funções de característica de maior peso associadas à etiqueta BEGIN . . 53
4.7 Funções de característica de maior peso associadas à etiqueta CONTINUE 53
4.8 Funções de maior peso associadas à etiqueta END 54
4.9 Funções de característica de maior peso associadas à etiqueta UNIQUE . 54
4.10 Funções de característica de maior peso associadas à etiqueta NEG 54
4.11 Resultados da avaliação do modelo para o GikiCLEF 2009 55
4.12 Classificação Linguística da WPT-05 60
4.13 Entidades extraídas com correspondências nas ontologias 61
4.14 Erros ortográficos em entidades extraídas 62

4.15 Exemplos de falta de artigos definidos em EG extraídas da WPT05 62
4.16 Exemplos de moradas extraídas da WPT-05 63
4.17 Entidades extraídas para os artigos da Wikipedia 64
4.18 Avaliação da Heurística 1 64
4.19 Avaliação da Heurística 2 65
4.20 Avaliação da Heurística 3 66
4.21 Referências extraídas para o artigo sobre Beja 67

Capítulo 1

Introdução

Os motores de pesquisa têm a função de recolher e organizar informação de forma a torná-la útil. Nos sistemas de recuperação de informação clássicos os documentos são organizados sob a forma de léxicos baseados no texto dos documentos, ignorando o significado semântico. Isto significa que um documento apenas pode ser encontrado com base no emparelhamento entre as palavras da consulta e as palavras que contém.

Porém, ao analisar um documento, conseguem-se identificar várias entidades (ex: pessoas, eventos, locais, temporais, etc) inferindo as relações entre elas e, recorrendo a uma base de conhecimento externa, depreende-se o conteúdo semântico (Berendt et al., 2002). Ao extrair e estruturar as entidades reconhecidas enriquece-se a informação disponível sobre um documento. Além de dados estatísticos sobre as palavras contidas, o documento passa a ter também meta-dados que descrevem a sua semântica. Isto permite a um utilizador procurar um documento tendo como base o contexto ou significado dos termos da consulta, em vez de apenas uma palavra ou sequência de palavras específicas.

Um dos elementos de contextualização presentes num documento são as suas referências geográficas. Uma análise da recolha da Web portuguesa de 2003 mostra que a informação geográfica está presente em páginas Web. De uma amostra de 32.000 documentos, obtidos a partir de uma recolha da Web portuguesa, Chaves and Santos (2006) mostram que 76% contêm uma localização geográfica . A existência de entidades geográficas é também frequente nas interrogações submetidas pelos utilizadores: Sanderson and Kohler (2004) ao analisarem os *logs* de cerca de 2 500 interrogações a um motor de pesquisa, verificaram que 18.6% continham um termo geográfico, e 14.8% um nome de um local o que leva a que a pesquisa de documentos tendo em conta a sua dimensão geográfica mereça um tratamento específico.

Aplicando técnicas de extracção de informação de texto para identificar entidades geográficas presentes num documento, e extraindo as relações entre elas, consegue-se criar um resumo geográfico. Ou seja, da mesma forma que numa biblioteca existe um catálogo que contém um resumo temático de cada livro, os documentos poderão estar associados a um espaço geográfico representado pelas relações entre as entidades geográficas presentes

no texto. O processo de gerar um resumo através das entidades reconhecidas é complexo devido à ambiguidade presente nos nomes, por exemplo, ao tentar compreender o âmbito geográfico de um documento, a partir do nome ”Camilo Castelo Branco”, um romancista português, poderá ser extraída a entidade geográfica ”Castelo Branco”, uma localidade em Portugal. Ou a expressão ”Odivelas”, poderá ser uma referência à freguesia pertencente ao Concelho de Ferreira do Alentejo ou à cidade com o mesmo nome no Distrito de Lisboa.

Através da análise de redes semânticas contendo os termos extraídos é possível resolver quase sempre o problema da ambiguidade. Depois de terem sido identificadas as expressões e palavras com referências geográficas, é necessário construir uma associação entre as referências encontradas e a área geográfica física a que se referem, aqui um dado endereço ou localização é associado a um identificador único, por exemplo uma coordenada geográfica ou uma chave de acesso num dicionário ou base de dados, ou um identificador numa ontologia geográfica.

Uma ontologia geográfica além de conter os nomes de entidades geográficas, descreve as relações entre os locais que representam. Por exemplo, um local poderá estar contido num outro local ou poderá ser também adjacente a um outro local. Uma ontologia geográfica é usada para identificar entidades geográficas extraídas e analisar as relações entre elas. O projecto GREASE (Geographic Reasoning for Search Engines) (Silva et al., 2006) investiga métodos de acesso a grandes colecções de documentos contendo textos e meta-dados com propriedades geográficas. No âmbito do projecto foram desenvolvidas duas ontologias geográficas, uma com âmbito no território português, a Geo-Net-PT e outra de carácter mundial, a World Geographic Ontology (WGO).

Os *Conditional Random Fields* (CRF) são uma teoria probabilística derivada das técnicas baseadas em *Hidden Markov Models* (HMM) usada na etiquetagem de dados sequenciais (Lafferty et al., 2001). Uma das áreas de aplicação é o reconhecimento de entidades em texto. Neste trabalho os CRF foram aplicados no reconhecimento de entidades geográficas mencionadas em textos escritos na língua portuguesa. As entidades identificadas são utilizadas para a geração de um resumo geográfico que consiste nas referências geográficas extraídas de um texto e resolvidas numa ontologia geográfica.

Esta dissertação decorreu no âmbito do projecto GREASE que investiga formas de capturar o conteúdo semântico dos documentos, por forma a melhorar as pesquisas sobre esses documentos, focando o interesse nas pesquisas geográficas. Uma das tarefas do GREASE consiste no desenvolvimento de técnicas de extracção de informação geográfica, isto é, reconhecimento automático de nomes geográficos, ou que indiciem localizações geográficas e a sua associação a localizações.

No âmbito de outras tarefas do projecto GREASE os resumos geográficos gerados serão usados para a computação de medidas de semelhança para uso em aplicações de recuperação de informação e visualização de informação geo-referenciada.

1.1 Objectivos

O objectivo principal desta dissertação foi o de desenvolver um processo para gerar resumos geográficos para documentos em português, o processo dividiu-se em 3 partes:

Reconhecer entidades geográficas num documento: Aplicando técnicas de prospecção de texto, em particular os CRF, extrair de um documento nomes de entidades com potencial significado geográfico como os nomes de ruas, concelhos, rios, serras, etc.

Desambiguar significados geográficos: Após terem sido extraídos os nomes de entidades geográficas é necessário decidir que significado estes têm. A mesma entidade pode ter diferentes significados, consoante o contexto onde se encontra. Analisando de uma forma global todas as entidades encontradas e utilizando uma base de conhecimento externa é possível desambiguar os possíveis significados geográficos de cada uma. Ou seja, eliminar referências com nomes idênticos aos extraídos. Por exemplo, ao extrair "Odivelas" de um texto juntamente com "Loures" ou "Lisboa" a probabilidade de que se refira a Odivelas como concelho é mais alta do que Odivelas como freguesia no concelho de Ferreira do Alentejo, a partir da informação existente na Geo-Net-PT, já que a primeira representa uma estrutura administrativa - concelho - mais importante e mais povoada.

Geração de um resumo geográfico: Um resumo geográfico é uma lista de entidades geográficas reconhecidas numa base de conhecimento externa, como uma ontologia. A representação do resumo gerado tem em consideração a sua utilização por outras aplicações, desta forma os resumos gerados são apresentados através de identificadores de conceitos associados numa ontologia, possibilitando a utilização explícita dos resumos por outras aplicações, com acesso à mesma ontologia.

A Figura 1.1 ilustra o processo seguido. As entidades geográficas são extraídas de um documento, é consultada uma base de conhecimento externo, sob a forma de uma ontologia geográfica de forma a associar-lhes o seu significado geográfico, é feito um processo de desambiguação e no final é gerado um resumo geográfico que representa o conteúdo geográfico do documento original.

1.2 Contribuições

O processo apresentado nesta dissertação é suportado pelo ambiente HENDRIX por mim desenvolvido para geração de resumos geográficos para documentos em português. O HENDRIX tem por base a Geo-Net-PT (Chaves et al., 2005), uma ontologia geográfica com âmbito no território português e um pacote de *software* de aprendizagem supervisionada, o Minorthird (Cohen, 2004) usado para extrair entidades geográficas de texto

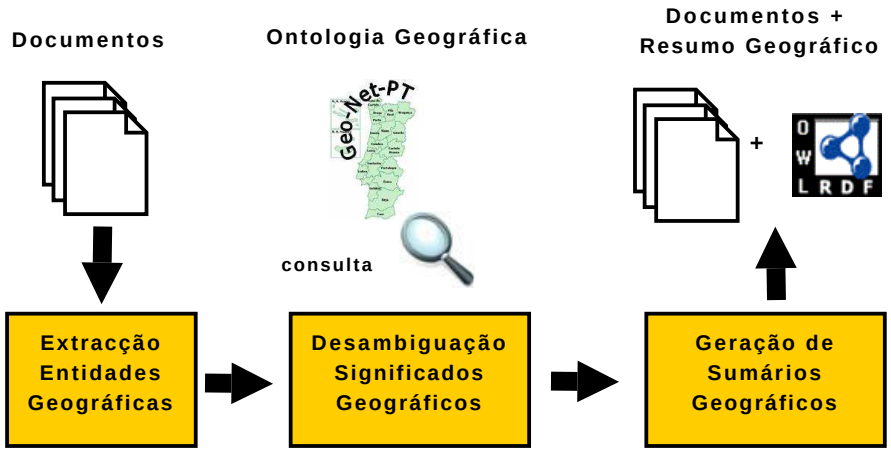

Figura 1.1: Processo de geração de resumos geográficos

com base na teoria dos CRF. Foi também desenvolvido o *software* PAREDES para fazer a desambiguação e classificação de entidades geográficas, com base na ontologia geográfica.

Os resumos gerados são apresentados em RDF (W3C, 2004), o formato de descrição de recursos da Web Semântica de maneira a poderem ser interpretados por outras aplicações (Berners-Lee et al., 2001).

O ambiente HENDRIX foi depois montado sobre um *cluster* de computadores formado com o Hadoop uma plataforma de *software* para computação distribuída que implementa o paradigma MapReduce (Dean and Ghemawat, 2004), de forma a poder efectuar o processo de extracção para grandes colecções de documentos, como recolhas da Web portuguesa.

A WPT-05, uma recolha da web portuguesa, foi a colecção de documentos usada para extrair as entidades geográficas. Foi necessário fazer a identificação linguística dos documentos que constituem. A versão em XML/RDF da WPT05 é agora distribuída com a indicação da língua na qual o documento se encontra escrito, mais o resumo geográfico de cada documento criado com o *software* desenvolvido nesta tese.

Dos cerca de 7,5 milhões documentos em português que fazem parte da WPT05, foram extraídas no total 78 326 entidades únicas. Para 18 586 (23.73%) foram encontradas correspondências em ontologias geográficas. A Tabela 1.1 apresenta uma síntese dos resultados obtidos. Os dados são disjuntos, o que significa que a mesma entidade poderá estar em mais do que uma ontologia.

Da extracção efectuada foram identificadas entidades com significado geográfico mas que não se encontram na ontologia geográfica usada, fortalecendo a necessidade de enriquecer essa mesma ontologia.

Ontologia	Nº Entidades	Percentagem
Geo-Net-PT 2.0	13 097	70.47%
World Geographic Ontology	2 191	11.79%
Wiki WGO 2009	8 742	47.04%

Tabela 1.1: Entidades extraídas com correspondências nas ontologias

1.3 Metodologia

O desenvolvimento do sistema proposto foi dividido em 5 tarefas:

- Análise de pacotes de *software* existentes para extracção de informação, com a aplicação da teoria de *Conditional Random Fields* no reconhecimento de entidades em textos.
- Análise dos dados na Geo-Net-PT, uma ontologia geográfico com âmbito no território português, contendo dados administrativos sobre distritos, concelhos, freguesias ruas, assim como suas as relações, dados populacionais e coordenadas geográficas.
- Desenvolvimento do *software* PAREDES para geração de resumos geográficos, com base nas entidades extraídas de documentos, e na Geo-Net-PT
- Geração dos resumos geográficos das páginas da WPT05 com o HENDRIX recorrendo a um *cluster* Hadoop.
- A geração dos resumos faz uso de várias heurísticas de avaliação. As heurísticas usadas são avaliadas com base em artigos da Wikipédia portuguesa referentes a capitais de Distrito.

1.4 Estrutura do documento

Esta dissertação encontra-se estruturada em 5 capítulos da seguinte forma: no Capítulo 2 é apresentado o trabalho relacionado, no Capítulo 3 é descrito o HENDRIX, no Capítulo 4 faz-se uma análise dos resultados obtidos, e no Capítulo 5 apresentam-se as conclusões e propostas de ideias para trabalho futuro.

Capítulo 2

Trabalho relacionado

O HENDRIX é um ambiente para extracção de informação geográfica e geração de resumos geográficos. Neste capítulo são descritos recursos e tecnologias usadas pelo HENDRIX. A extracção de informação geográfica de documentos é feita recorrendo a uma técnica de *Text Mining* baseada em aprendizagem supervisionada (Mitchell, 1997). A extracção poderá ser efectuada em larga escala recorrendo ao Hadoop, uma plataforma para utilização de *clusters* para processamento distribuído com base no paradigma MapReduce (Dean and Ghemawat, 2004). Os resumos geográficos são gerados tendo por base ontologias geográficas. É introduzido o evento HAREM (Mota and Santos, 2008a), onde diversos sistemas de reconhecimento de entidades são avaliados em conjunto. As ontologias geográficas usadas pelo HENDRIX são descritas, assim como o modelo de dados usado na sua construção. É apresentada também, na primeira secção, a terminologia usada ao longo desta dissertação.

2.1 Terminologia

A terminologia usada ao longo desta dissertação é baseada na terminologia apresentada por Chaves (2009) e pretende ser uma sistematização da terminologia a usar em Recuperação de Informação Geográfica (RIG) e Extracção de Informação Geográfica (EIG).

Entidade Extraída (EX): uma expressão ou conjunto de palavras extraídas de um texto.

Tipo de Conceito Geográfico (TCG): conceito geográfico associado a uma entidade, por exemplo: país, cidade, avenida, rua, distrito, concelho, freguesia. No âmbito da Geo-Net-PT é denominado *feature type*

Referência Geográfica (RG) : uma entidade geográfica definida sem ambiguidade, por um identificador único numa ontologia geográfica. Denominado como *feature* no âmbito da Geo-Net-PT.

Atomização: O processo no qual os constituintes de um texto (palavras, sinais de pontuação) são divididos. O resultado são as unidades mínimas que constituem um texto, denominados termos. Há várias técnicas para conseguir isto, uma deles consiste em usar uma expressão regular.

Expressão Regular : Permite identificar cadeias de caracteres de forma concisa e flexível sem precisar listar todos os elementos do conjunto.

Termo: A unidade mínima que constitui um texto, pode ser uma palavra, um sinal de pontuação, também denominado por átomo ou *token*.

2.2 Text Mining

A linguagem natural foi desenvolvida para comunicação entre humanos e consequentemente para ser interpretada por humanos. A compreensão de qualquer documento textual por uma máquina é impraticável. No entanto é possível extrair pequenas quantidades de informação útil dos documentos seguindo determinados padrões.

A disciplina de *Text Mining* estuda métodos para extracção de informação de textos. Existem várias estratégias para conseguir extrair expressões ou palavras ao processar um texto. Uma delas consiste em definir regras de reconhecimento, um método linguístico. Estas detectam provas que impliquem a presença de entidades a capturar. As regras são codificadas à mão e normalmente têm como base a gramática da língua na qual os documentos se encontram escritos.

Outra estratégia consiste na aprendizagem automática dos padrões a extrair, recorrendo a métodos de aprendizagem supervisionada. Não são necessárias regras complexas, mas é necessário uma colecção de documentos anotados. Numa primeira fase o algoritmo, através do cálculo de probabilidades, aprende a detectar padrões no texto de modo a inferir as suas próprias regras para detectar a presença de uma entidade. Numa segunda fase, as regras inferidas são aplicadas a um texto não anotado, sendo usadas para calcular a probabilidade de uma dada palavra ser ou não uma entidade a extrair, ou parte de uma expressão a extrair. Podem também ser aplicadas a um outro texto anotado de forma a medir o desempenho das regras geradas.

2.2.1 Pré-processamento

De forma a que documentos possam ser processados automaticamente é necessário em geral haver previamente uma transformação do documento para uma representação textual estruturada. A esta fase chama-se pré-processamento.

O método de pré-processamento aplicado a um documento está relacionado com o objectivo a alcançar com o processamento do texto. Segundo Feldman (2006) a fase de

pré-processamento divide-se em três classes: processamento preparatório, processamento de linguagem natural e processamento considerando o domínio do problema.

A classe de processamento preparatório trata de transformar dados de uma representação não textual, por exemplo voz, documentos obtidos através do reconhecimento óptico de caracteres, documentos partes de recolhas da World Wide Web contendo etiquetas HTML ou documentos PDF, numa representação textual que possa ser processada por *software*.

O processamento de linguagem natural, analisa um texto com base na gramática da língua em que se encontra escrito. A atomização do texto poderá ser feita de acordo com o seu significado morfológico, cada palavra do documento é analisada e classificada morfologicamente. O resultado deste pré-processamento poderá ser usado noutras tarefas.

O processamento considerando o domínio do problema, tem como objectivo apresentar o significado que o documento tem no problema em questão. poderá usar o resultados do processamento das outras duas classes. Tipicamente, nesta fase são aplicadas técnicas de categorização e de extracção de informação, a seguir descritas.

2.2.2 Categorização e Extracção de Informação

A categorização tem como objectivo atribuir uma categoria, um conjunto de conceitos ou palavras chave, a um documento. O conjunto de categorias é normalmente determinado manualmente, é fechado e relativamente pequeno.

A Extracção de Informação (EI) extrai a informação relevante de um documento, com base em padrões, e apresenta-a de uma forma estruturada. É definida por Cunningham (2005) como o processo que recebe um documento de entrada e produz como saída dados não ambíguos para servirem um propósito definido, como por exemplo, serem apresentados a um utilizador, armazenados numa base de dados ou servirem o processo de indexação na área Recuperação de Informação (RI).

O processo de RI apenas encontra documentos e apresenta-os ao utilizador, enquanto que uma aplicação de EI analisa um documento e apresenta apenas a informação para o qual o utilizador está interessado ou para um propósito em questão.

Uma das tarefa relacionadas com a extracção de informação é a identificação de determinado tipo de entidades num texto. Designa-se por Reconhecimento de Entidades Mencionadas (REM) a tarefa de identificação de Entidades Mencionadas (EM) presentes num texto e da interpretação do seu significado semântico. Constituem exemplos de EM referências a nomes de pessoas, organizações, ou locais. O REM pode ser dividido em duas sub-tarefas:

Identificação: selecciona os termos que compõem uma EM.

Classificação: determina propriedades linguísticas das EM, como por exemplo o seu significado semântico ou a sua morfologia.

2.2.3 Extracção de Informação Geográfica

A tarefa de extracção de informação geográfica de textos está dividida em duas sub-tarefas. Densham and Reid (2003) definem os termos *geoparsing* e *geocoding*. Estas duas sub-tarefas têm o mesmo propósito que os processos de identificação e classificação da tarefa genérica de REM, mas num contexto geográfico.

À primeira sub-tarefa de identificação de referências geográficas dá-se o nome de geo-reconhecimento (*geoparsing*), que consiste em extrair do texto palavras ou expressões que indiciem referências geográficas. Aqui é importante desambiguar o que poderão ser, ou não, e dependente do contexto, os nomes de entidades geográficas.

Numa segunda fase, depois de terem sido identificadas as palavras e expressões com referências geográficas, é necessário classificá-las segundo o seu significado geográfico. A este processo dá-se o nome de geo-codificação (*geocoding*) ou geo-classificação. Atra–vés da comparação dos termos extraídos com os nomes num léxico geográfico ou numa ontologia geográfica eliminam-se os falsos positivos – termos extraídos mas que na realidade não correspondem a nenhum conceito geográfico – e associam-se os que fazem parte de um dicionário de nomes a um ou mais conceitos geográficos. Outra hipótese a considerar, é a extracção de termos que efectivamente têm algum significado geográfico, mas que não fazem parte do dicionário de nomes usado.

A utilização de uma ontologia geográfica tem vantagens em relação aos léxicos de nomes uma vez que contem além dos nomes e conceitos geográficos que estes representam, as relações entre os vários conceitos.

2.2.4 Desambiguação

As referências geográficas identificadas nos textos não são suficientes para determinar um conceito geográfico único associado à palavra extraída do texto. O mesmo nome pode ser usado para mencionar mais do que um local. Ao ser extraído um termo candidato a referência geográfica, por exemplo ”Souto”, este pode ser referente às seguintes entidades do domínio administrativo, na Geo-Net-PT:

- aldeia da freguesia de Pombal
- freguesia no concelho de Abrantes
- freguesia no concelho de Arcos de Valdevez
- freguesia no concelho de Penedono
- freguesia no concelho do Sabugal
- freguesia no concelho de Santa Maria da Feira

- freguesia no concelho de Terras de Bouro

Além disso Souto é uma categoria do domínio físico, pode representar uma mata de castanheiros. Por isso, é necessário a seguir à fase de identificação, iniciar o processo de desambiguação, que consiste em seleccionar destas referências aquela que estará a ser mencionada. Três diferentes tipos de ambiguidade podem ocorrer: ambiguidade no referente, ambiguidade na referência, e ambiguidade na classe do referente. O resto desta secção apresenta cada um em detalhe.

1. Ambiguidade no referente

O mesmo nome pode ser usado para mencionar mais do que um local, ou seja, a mesma entidade pode ter diferentes significados geográficos, consoante o contexto onde se encontra. Este caso poderá ser desambiguado recorrendo às seguintes heurísticas.

Um referente por documento: quando uma referência geográfica ambígua é usada várias vezes no mesmo documento, é provável que se refira a apenas um dos seus possíveis significados, daí que o seu significado poderá ser desambiguado assumindo que todas as ocorrências dessa referência têm o mesmo significado (Gale et al., 1992).

Referências geográficas relacionadas: como foi referido acima, referências geográficas dentro do mesmo documento tendem a referir localidades relacionadas. Esta relação poderá ser estabelecida através de propriedades geo-espaciais, como a proximidade, ou topológicas, definidas através de uma ontologia. Rauch et al. (2003) mostram que há uma correlação geo-espacial alta entre entidades geográficas que estão próximas num texto.

Significado mais comum: as localidades mais importantes são provavelmente mais referenciadas. Por exemplo, é mais provável que o termo ”Lisboa” seja uma referência à cidade do que a uma rua ou praça com o mesmo nome, se não estiver precedido de ”Rua de” ou ”Praça de”. A importância de um local também pode ser estimada através de dados demográficos, por exemplo os locais com maior população são mais importantes, ou através de níveis hierárquicos, as cidades têm mais importância que aldeias.

2. Ambiguidade na referência

O mesmo local pode ser mencionado através de mais do que um nome. Alguns locais são referidos por nomes distintos dos oficiais, atribuidos pelas entidades que fazem a gestão de dados administrativos geográficos. Por exemplo, ”Baixa de Lisboa” ou ”Baixa Pombalina” são expressões usados para referir uma zona específica no centro da cidade de Lisboa. Expressões como estas podem ocorrer no texto, é necessário adicionar dados

Nome próprio	Nome de local
Césaria Évora	Évora
Almada Negreiros	Almada
Camilo Castelo Branco	Castelo Branco
Salgueiro Maia	Maia

Tabela 2.1: Apelidos correspondentes a nomes de locais

extra, nomes alternativos, à base de conhecimento externa usada para que se possa com sucesso classificar geograficamente a expressão ou nome extraído. Outra alternativa é usar mais do que uma base de conhecimento externa.

3. Ambiguidade na classe do referente

O mesmo nome poderá representar outras classes de entidades. Há contextos em que o nome de uma localidade pode ser usado com outro significado, como por exemplo, no caso dos municípios. Onde, o mesmo nome pode referir-se a uma pessoa ou ao nome de uma companhia. Portanto, o significado que o nome pode ter varia consoante o contexto. Por exemplo, muitos nomes de locais portugueses são também apelidos, alguns de personalidades portuguesas, a Tabela 2.1 mostra alguns exemplos.

2.2.5 Referências implícitas

Existem outras entidades presentes em textos que, mesmo não sendo nomes explícitos de lugares ou localizações geográficas, têm um âmbito geográfico associado, como aeroportos, estádios, edifícios históricos, monumentos, nomes de organizações ou acontecimentos. Um acontecimento está normalmente associado a um espaço geográfico, e uma organização, seja uma empresa ou uma instituição, tem uma sede num local ou representantes em vários locais. Estas referências implícitas podem ser também usadas no processo de desambiguação e enriquecer os resumos geográficos gerados (Cardoso et al., 2008).

2.2.6 Conditional Random Fields

Os *Conditional Random Fields* (CRF) (Lafferty et al., 2001) são um modelo probabilístico para calcular a probabilidade de cada uma das possíveis etiquetas de classificação, dada um sequência de observações. É um modelo derivado dos *Hidden Markov Models* (HMM) (Rabiner, 1989), e tem como vantagem o facto de ser menos restrito em relação a pressupostos de independência entre as variáveis de observação. Este tipo de modelos são aplicados em problemas de etiquetagem de sequências estruturadas, tal como o texto em linguagem natural.

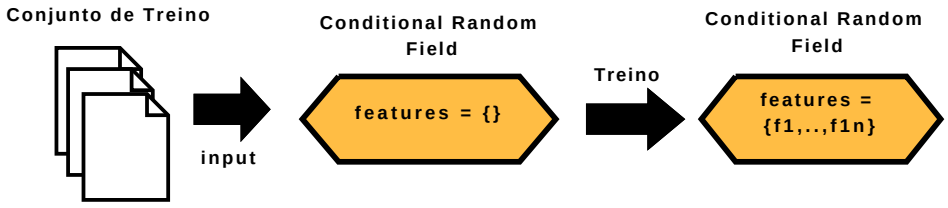

Figura 2.1: Treino de Conditional Random Fields

Os CRF permitem calcular $p(\vec{y}|\vec{x})$ de um resultado de saída $\vec{y}$ dado um conjunto de observações de entrada $\vec{x}$. A dependência condicional de cada y_i em $\vec{x}$ é especificado por um vector definido da forma $f = (f_1, f_2, ..., f_m)$ de funções característica (*features*) e um vector de pesos da forma $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)$. As funções característica geradas são da forma $f(i, y_{i-1}, y_i, \vec{x})$, ou seja podem depender do valor da etiqueta anterior na parte da sequência de entrada já classificada, da posição na sequência, e de toda a sequência de entrada.

O modelo atribui a cada função característica um peso e combina-as para determinar a probabilidade de cada y_i. Outra vantagem dos CRF sobre os HMM prende-se com o facto de poderem conter um número arbitrário de funções característica que podem avaliar toda a sequência de entrada a qualquer altura durante o processo de treino.

Klinger and Tomanek (2007) explicam a construção de modelos clássicos probabilísticos, assim como uma descrição detalhada da construção de um modelo baseado em CRF e de que forma este se relaciona com outros modelos, como *Naïve Bayes*, HMM e *Maximum Entropy Models* (MEM).

Há três problemas clássicos relacionados com os CRF:

Treino: Dado um conjunto de dados de treino $(\vec{x}, \vec{y})$, onde $\vec{x}$ representa a sequência de entrada, e $\vec{y}$ as etiquetas dadas a cada elemento de $\vec{x}$ encontrar os parâmetros λ do CRF que vão maximizar a verosimilhança com os dados de treino.

Classificação: Dado um CRF de parâmetros λ e uma sequência $\vec{x}$, encontrar a etiqueta mais provável $y = argmax_y\ p_\lambda(y|x_i)$.

Avaliação: Dado um CRF de parâmetros λ, uma sequência $\vec{x}$ e uma sequência de etiquetas $\vec{y}$, encontrar a probabilidade condicional $p_\lambda(\vec{y}|\vec{x})$

Destes interessam para o trabalho desenvolvido dois, o Treino e a Classificação. Na fase de Treino, é gerado o modelo com base nas base nas características extraídas dos documentos anotados, como mostra a Figura 2.1. Na Classificação o modelo gerado é aplicado a um documento não anotado, atribuindo com base nas funções de característica geradas etiquetas a cada palavra do documento. De forma a poder avaliar o modelo gerado este é aplicado noutro conjunto de documentos também anotado, mas disjunto do

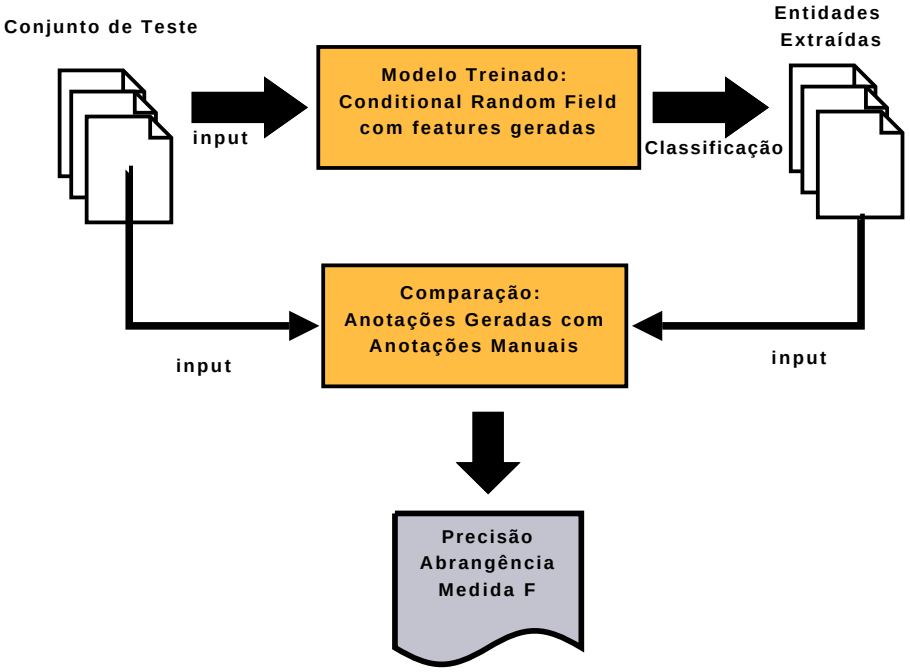

Figura 2.2: Classificação usando um Conditional Random Field

primeiro. As entidades extraídas são depois comparadas com as entidades anotadas, como exemplificado na Figura 2.2.

2.2.7 Medidas de Avaliação

A qualidade do modelo gerado pode ser avaliado usando as medidas de Precisão, Abrangência e a Medida F, definidas da seguinte maneira:

- **Precisão** : mede a ”qualidade” de resposta do sistema, ou seja calcula a proporção de entidades extraídas correctamente, todas as entidades que de facto são entidades geográficas, em relação a todas as entidades extraídas.

$$\text{Precisão} = \frac{(entidades\ geograficas\ extraidas) \cap (entidades\ extraidas)}{(entidades\ extraidas)}$$

- **Abrangência**: mede a ”quantidade” de respostas correctas dadas, calcula a proporção de entidades geográficas extraídas em relação ao universo de possíveis entidades com significado geográfico.

$$\text{Abrangência} = \frac{(entidades\ geograficas\ extraidas) \cap (entidades\ extraidas)}{(entidades\ geograficas\ presentes\ num\ documento)}$$

- **Medida F**: combina as métricas de precisão e de abrangência de acordo com a seguinte fórmula:

$$\text{Medida F} = \frac{2 \times precisao \times abrangencia}{precisao + abrangencia}$$

Software	Formato Entrada	Formato Saída	Linguagem	UI
Sunita Sarawagi's CRF	Texto formatado	Texto formatado	Java	Não
MALLET	Texto formatado	Texto formatado	Java	Não
Minorthird	XML	Vários	Java	Sim

Tabela 2.2: Algumas propriedades do *software* analisado

2.2.8 Software Analisado

Foram analisados três pacotes de *software* que aplicam a teoria dos CRF ao reconhecimento de entidades em texto, na Tabela 2.2 são apresentadas algumas das suas características.

Sunita Sarawagi's CRF package: é uma implementação bastante simples e resume-se apenas à utilização de CRF para etiquetagem de textos. Há documentação e alguns tutoriais sobre como reutilizar o código. Por outro lado, o texto anotado para aprendizagem tem que seguir um formato bastante específico. O ficheiro com os dados de entrada tem que ter um *token* por linha com a etiqueta correspondente a esse *token* na mesma linha separados por um carácter especial. Todos os documentos são passados num único ficheiro, sendo os documentos separados no ficheiro por uma linha em branco. O documento de saída é também apresentado no mesmo formato. Um único ficheiro com o termo ou termos identificados seguidos de um separador e a etiqueta associada. Tendo sido criada com o intuito de ser usada noutras aplicações, existem alguns trabalhos desenvolvidos que o usam como base. Jungermann (2006) usa este pacote de *software* para criar um *plugin* baseado em CRF para o RapidMiner (antigo YALE), um ambiente para experiências de aprendizagem supervisionada e prospecção de dados. É distribuído sob a University of Illinois/NCSA Open Source License: http://www.otm.illinois.edu/faculty/forms/opensource.asp. Está disponível em: http://crf.sourceforge.net

MALLET: aplica técnicas de aprendizagem automática para processamento de linguagem, classificação de documentos, *clustering*, extracção de informação (McCallum, 2002). Inclui métodos para classificação de documentos como *Naïve Bayes, Maximum Entropy*, Árvores de Decisão e código para avaliar o desempenho, usando as métricas definidas na seccção anterior. Além de classificação permite também a etiquetagem sequencial para aplicações de extracção de entidades de texto, usando para isso algoritmos baseados em HMM, MEM e CRF. O formato dos dados é semelhante ao do *Sunita Sarawagi's CRF package*. É distribuído sob a Common Public License Version 1.0 (CPL): http://www.opensource.org/licenses/cpl1.0.php. Está disponível em: http://mallet.cs.umass.edu

Minorthird: permite fazer categorização e extracção de informação de textos. Contém

```
Alista-se com 25 anos na armada que foi a <local>India</local>,
comandada por <nome>Francisco de Almeida</nome>, embora o seu
nome nao figure nas cronicas; sabe-se no entanto que ali permaneceu
oito anos, que esteve em <local>Goa</local>, <local>Cochim</local>,
<local>Quiloa</local>,que acompanhou <nome>Diogo Lopes de Sequeira
</nome> a <local>Malaca</local>, viagem que acabou em naufragio ....
<nome>Fernao de Magalhaes<nome> morre nas <local>Filipinas</local>
no curso daquela expedicao, posteriormente chefiada por <nome>Juan
Sebastian Elcano<nome> em 1522</EM>
```

Figura 2.3: Exemplo de um texto de entrada para o Minorthird

além dos CRF outros modelos de extracção e classificação de documentos, como *Support Vector Machines*, MEM, *Decision Trees*, *Clustering* (Cohen, 2004). O formato dos dados de entrada é mais flexível, bastando apenas recorrer a etiquetas para marcar as entidades a serem aprendidas a reconhecer, como mostra a Figura 2.3, o que é uma vantagem em relação aos outros pacotes de *software* analisados. Tem também uma interface gráfica, que torna mais fácil a sua utilização. É distribuído sob a BSD License: http://www.opensource.org/licenses/bsd-license.php. Está disponível em: http://minorthird.sourceforge.net

O Minorthird foi o pacote escolhido. O código do *Sunita Sarawagi's CRF package* é usado no próprio Minorthird para aplicar a teoria dos CRF a reconhecimento de entidades em textos. Alguns dos factores que fizeram a escolha cair sobre o Minorthird foram:

Formato dos dados: para a fase de treino basta marcar as entidades a aprender com etiquetas XML, como na Figura 2.3 Na fase de Classificação pode-se escolher como os dados são apresentados, também com etiquetas nas entidades reconhecidas, ou apenas um ficheiro de saída identificado o documento e as entidades reconhecidas.

Tutorais: estão disponíveis uma série de materiais de apoio que permitem explorar de forma interactiva as funcionalidades da aplicação. Estes foram úteis para perceber o seu funcionamento.

Comunicação com Autores: em diversas ocasiões foi possível comunicar com os autores do Minorthird para ajuda e esclarecimento de dúvidas.

Código Java aberto: permitiu uma fácil integração das funcionalidades do Minorthird no HENDRIX, e consequentemente no cluster formado pelo Hadoop.

No entanto, muitos outros pacotes existem, e à medida que o trabalho foi sendo desenvolvido novos pacotes surgiram com suporte para CRF, como os listados na Tabela 2.3

Software	Linguagem	URL
CRFSuite	C	http://www.chokkan.org/software/crfsuite/
Xcrf	Java	http://treecrf.gforge.inria.fr/
CRF++	C++	http://crfpp.sourceforge.net/
FlexCRFs	C++	http://flexcrfs.sourceforge.net/
JProGraM	Java	http://www.dii.unisi.it/ freno/JProGraM.html
Lingpipe	Java	http://alias-i.com/lingpipe/

Tabela 2.3: Outros pacotes de *software* com suporte para CRF

Minorthird

O Minorthird faz o reconhecimento de entidades em texto ou extracção através de classificação dos termos. Começa por gerar *features* para todos os termos contidos num texto. Por exemplo, tendo em consideração a seguinte frase:

"A viagem de **Fernão Magalhães** é relatada no diário de **Antonio Pigafetta**."

numa tarefa de reconhecimento de entidades mencionadas para nomes de pessoas, na frase acima os termos a *bold* marcam as entidades a aprender a extrair. O objectivo é gerar um modelo, a partir do exemplo dado, que extrai nomes de pessoas de texto. A frase é atomizada e cada termo é representado segundo uma classe e um conjunto de *features*, como mostra a Tabela 2.4.

Os termos marcados como POS apresentam as entidades que se querem aprender a extrair e NEG entidades a não extrair. Os termos POS são ainda classificados em quatro categorias:

- **BEGIN**: um termo inicial de uma entidade a ser extraída
- **END**: um termo final de uma entidade a ser extraída
- **CONTINUE**: um termo que faz parte de uma entidade a ser extraída, e não é o inicial nem o final
- **UNIQUE**: um único termo que constituí a entidade a ser extraída

Às *features* geradas para cada termo é associado um peso. O modelo gerado é assim constituído pelas *features* extraídas a partir de um texto anotado e o seu peso. Este modelo é depois gravado como um objecto Java, denominado *Annotator*. As *features* geradas para cada umas das quatro categorias podem utilizar as seguintes propriedades do texto e das anotações:

- número de termos à esquerda ou direita da entidade marcada, dos quais se podem considerar *features* (por omissão o valor é 3).

Entidade/Termo	**Classe**	***Features* extraídas**
A	NEG	tokens.eq.charTypePattern.x+=0; previousLabel.1.null=1;
viagem	NEG	tokens.eq.charTypePattern.x+=1; previousLabel.BEGIN=0;
de	NEG	tokens.eq.charTypePattern.x+=1; previousLabel.NEG=1;
Fernão	POS	charTypePattern.X+x+=1; previousLabel.1.NEG=1;
Magalhães	POS	right.tokenNeg_0.eq.charTypePattern.X+x=0; previousLabel.1.BEGIN=1;
é	NEG	tokens.eq.charTypePattern.X+=0; previousLabel.1.END=1;
relatada	NEG	tokens.eq.charTypePattern.x+=1; previousLabel.1.END=0;
no	NEG	tokens.eq.charTypePattern.x+=1; previousLabel.1.END=0;
diário	NEG	tokens.eq.charTypePattern.x+=1; previousLabel.1.END=0;
de	NEG	tokens.eq.lc.de=1; right.token_0.eq.charTypePattern.X+x=0;
António	POS	tokens.eq.charTypePattern.X+=1 previousLabel.1.END=0;
Pigafetta	POS	tokens.eq.charTypePattern.X+=1; previousLabel.1.BEGIN=1;

Tabela 2.4: Exemplos de *features* geradas

- classificação do termo anterior: BEGIN, END, UNIQUE, CONTINUE ou NEG
- padrão que caracteriza o termo anterior, por exemplo:
  - charTypePattern.X+x+: o termo começa com um carácter maiúsculo
  - charTypePattern.X+: o termo contem apenas caracteres maiúsculos
  - charTypePattern.x+: o termo contem apenas caracteres minúsculos
- termo actual

O modelo gerado pode ser aplicado depois a textos não anotados para extrair entidades. Um texto do qual se querem extrair entidades é atomizado e são utilizadas as funções de característica (*features*) do modelo para calcular a probabilidade de cada termo ter uma das cinco classificações, BEGIN, END, UNIQUE ou CONTINUE caso faça parte de uma entidade a extrair ou NEG para uma entidade a não extrair. Um termo é classificado segundo a etiqueta com a probabilidade mais alta.

De forma a conseguir gerar um modelo para extracção de entidades é necessário haver previamente um texto anotado, para que se possam gerar as *features* que vão constituir o modelo de extracção. Na próxima secção apresenta-se o HAREM, o evento para avaliação de sistemas de reconhecimento de entidades. Um dos artefactos disponíveis nesse evento é uma colecção de textos em português anotada, denominada Colecção Dourada, utilizada neste trabalho precisamente para treinar o processo de EI com CRF.

2.3 HAREM - Avaliação de Reconhecimento de Entidades Mencionadas

O HAREM é um evento de avaliação conjunta em reconhecimento de entidades mencionadas para o português (Mota and Santos, 2008a) (Santos and Cardoso, 2008), criado e organizado pela Linguateca (Santos, 2009).

Num modelo de avaliação conjunta vários grupos comparam o desempenho dos seus sistemas, usando para isso um conjunto de recursos em comum, e uma métrica consensual. A metodologia do HAREM inclui a definição das directivas de etiquetagem dos textos, a especificação das tarefas de avaliação e o processo de criação das colecções de texto. Além de usar documentos em português, destaca-se de outros eventos em vários aspectos, dos quais saliento pela sua importância para esta dissertação:

Colecções com diversos tipos de texto: Existem diferenças significativas no teor e na distribuição de EM entre géneros textuais. Uma vez que os sistemas de REM participantes podem ter sido desenvolvidos para processar diferentes tipos de texto, as colecções usadas contêm textos de vários géneros textuais e de várias variantes de português. A recolha da web portuguesa, WPT05, contém vários géneros textuais, e variantes do português, utilizando esta colecção para o processo de treino, o modelo fica mais adaptado à colecção de documentos onde vai ser aplicado.

Avaliação independente das tarefas: As tarefas de identificação e de classificação são avaliadas em separado, para diagnosticar detalhadamente o desempenho dos sistemas. É possível desta forma avaliar apenas a identificação sendo a classificação e a sua avaliação feita com as ontologias geográficas.

Avaliação selectiva: A avaliação adapta-se às características de cada sistema, medindo o desempenho das saídas segundo um sub-conjunto de categorias e tipos de EM pré-seleccionados pelo sistema participante. É possível desta forma avaliar o sistema apenas para a identificação de entidades geográficas.

Anotação em contexto: A anotação manual das colecções tem em consideração o contexto onde se insere a EM, e a classificação é feita atendendo a critérios semânticos. Por exemplo um nome de um jornal poderá ser classificado como uma organização (uma empresa), um local (de publicação) ou uma pessoa (um entrevistador) dependendo do contexto. Isto permite um maior rigor na avaliação do modelo de CRF gerado.

As directivas de etiquetagem do HAREM são seguidas pelos participantes no desenvolvimento dos sistemas, e são usadas na anotação manual da colecção de textos. A categorização é composta por uma hierarquia de dois níveis, denominados categorias e

tipos. As categorias representam as classes semânticas principais das EM e são compostas por vários tipos, que são especializações de cada categoria. Cada tipo pertence a uma única categoria apenas, e cada EM é classificada por uma categoria e por um tipo, no mínimo. Foram definidas 10 categorias e 41 tipos. As 10 categorias com os correspondentes tipos são apresentados na Tabela 2.5.

Para as categorias "LOCAL" e "TEMPO" foi ainda definido mais um nível hierárquico, sub-tipo. Um LOCAL do tipo FISICO, poderá ter como sub-tipo: ILHA, AGUACURSO, PLANETA, REGIAO, RELEVO, AGUAMASSA, OUTRO. Um local HUMANO: RUA, PAIS, DIVISAO, REGIAO, CONSTRUCAO, OUTRO. Um local VIRTUAL: COMSOCIAL, SITIO, OBRA, OUTRO. Uma entidade TEMPO, com tipo TEMPO_CALEND poderá ter como sub-tipo: HORA, INTERVALO, DATA, OUTRO;

2.3.1 Colecção Dourada

A Colecção Dourada (CD) de cada evento HAREM consiste num conjunto de textos marcados com as EM identificadas e classificadas correctamente por todos os participantes segundo o conjunto de directivas de Mota and Santos (2008b) e usando as etiquetas apresentadas na Tabela 2.5. As classificações categorizam cada EM a ser identificada pelos sistemas. A CD é usada para avaliar os sistemas participantes, comparando a CD original com as anotações produzidas pelos sistemas participantes.

Decorreram até ao momento 3 eventos HAREM: Primeiro HAREM (2005), MiniHAREM (2006), Segundo HAREM (2008), de onde resultaram 3 CD constituídas por vários géneros textuais:

Web: textos extraídos de páginas HTML da recolha da *web* portuguesa WPT-03 e da recolha da *web* brasileira WBR-99.

Jornalístico: textos retirados de corpora jornalísticos CETEMPúblico, CETENFolha, Avante!, Viseu Diário, Diário do Minho e Jornal de Macau.

Entrevista: textos transcritos de entrevistas orais cedidas pelo Museu Pessoa de Portugal e do Brasil.

Técnico: textos técnicos e científicos extraídos a partir de relatórios contidos no WPT03 e tratados no Corpógrafo.

Correio Electrónico: excertos de mensagens da *mailing list* brasileira ANCIB (www.ancib.org.br), e do corpus de mensagens CONE(www.linguateca.pt/corpora_info.html).

Expositivo: textos retirados de várias fontes de informação da web, como a Wikipedia (pt.wikipedia.org).

Categoria	Tipo
ABSTRACCAO	DISCIPLINA ESTADO IDEIA NOME OUTRO
ACONTECIMENTO	EFEMERIDE EVENTO ORGANIZADO OUTRO
COISA	CLASSE MEMBROCLASSE OBJECTO SUBSTANCIA OUTRO
LOCAL	FISICO HUMANO VIRTUAL
OBRA	ARTE PLANO REPRODUZIDA OUTRO
ORGANIZACAO	ADMINISTRACAO EMPRESA INSTITUICAO OUTRO
PESSOA	CARGO GRUPOCARGO GRUPOIND GRUPOMEMBRO INDIVIDUAL MEMBRO POVO OUTRO
TEMPO	DURACAO FREQUENCIA GENERICO TEMPO_CALEND OUTRO
VALOR	CLASSIFICACAO MOEDA QUANTIDADE OUTRO
OUTRO	

Tabela 2.5: Categorias e tipos definidos no segundo HAREM

Variante	CD de 2005	CD de 2006	CD de 2008
Portugal	38.472 (41,44%)	29.864 (47,81%)	44.555 (59,93%)
Brasil	49.737 (53,58%)	32.597 (52,19%)	29.795 (40,07%)
África	1.435 (1,55%)	-	-
Ásia	3.186 (3,43%)	-	-

Tabela 2.6: Distribuição de termos segundo a variante de português.

Literário: extractos de obras literárias de diversos autores portugueses, brasileiros, angolanos e moçambicanos.

Político: extractos dos corpora EuroParl (people.csail.mit.edu/koehn/publications/europarl/), ECI-EBR (www.linguateca.pt/corpora_info.html) e de discursos de origem timorense.

Os documentos que fazem parte da CD abrangem as variantes do português, na Tabela 2.6 é apresentada a distribuição de termos segundo a variante de português do documento de onde foram retirados.

A CD pretende representar o que a comunidade entende ser o resultado ideal da tarefa de REM, mas as anotações estão longe de representar o que se espera que os sistemas de REM actuais consigam realizar. Durante o processo de anotação das CD foi frequente encontrar diferentes interpretações no sentido semântico de várias EM por parte dos anotadores, e leituras diferentes do âmbito semântico dado pela categorização HAREM, o que mostra que há um limite para os desempenhos da tarefa de REM imposto pela própria ambiguidade da língua. Uma vez que até os humanos discordam entre si na marcação de certas EM, não faz sentido exigir aos sistemas de REM que consigam marcar correctamente as EM nesses casos.

2.4 Ontologias Geográficas

Uma ontologia é uma descrição formal de conceitos e das relações entre eles, dentro de um domínio específico. Representa o conhecimento humano de forma a ser interpretado por uma máquina. Um ontologia geográfica, ou uma geo-ontologia, é uma ontologia para o domínio geográfico. Os conceitos geográficos – cidades, ruas, concelhos – e a forma como se relacionam – uma freguesia é parte de um concelho, e este parte de um distrito – são descritos formalmente.

Conceptualmente, uma ontologia poderá ser vista como um grafo onde os nós representam conceitos geográficos e as arestas relações entre os conceitos. Os nós podem ter propriedades como população, ou coordenadas geográficas. As arestas representam relações, como por exemplo *parte-de* ou *adjacente-a*, para indicar que um conceito ge-

ográfico é parte de outro, ou que são geograficamente adjacentes. Nesta secção é apresentado um modelo para construção de ontologias geográficas, integrando dados de diferentes fontes. São também apresentadas duas ontologias construídas com base nesse mesmo modelo. Estas são usadas na fase de classificação das entidades entidades geográficas identificadas, são também usadas para extrair relações entre as entidades classificadas.

2.4.1 Geographic Knowledge Base

No âmbito do projecto GREASE foi desenvolvido um sistema de informação para construção de ontologias geográficas. O Geographic Knowledge Base (GKB) (Chaves et al., 2007) foi desenho com o objectivo construir um repositório para integração de dados vindos de diferentes fontes, dentro de um esquema comum, de forma a ser utilizado por aplicações que usam técnicas de extracção de informação na prospecção de conceitos geográficos, tendo em conta o seu âmbito semântico. Os dados carregados são organizados em modelos de informação, cada um representando um conjunto de *features* geográficas relacionadas.O GKB organiza a informação segundo domínios, cada domínio pode organizar a informação segundo o modelo descrito em seguida.

O meta-modelo base da versão actual do GKB (2.1) é apresentada na Figura 2.4. A classe *Feature* é associada à classe *Type* que guarda os tipos, por exemplo uma *feature* - que representa um conceito geográfico único - com o nome ”Liberdade” é do tipo ”Avenida”. A classe *Type-Relationship* guarda as relações entre tipos, por exemplo um município é parte de um país. A classe *Relationship-Type* guarda as relações entre con-

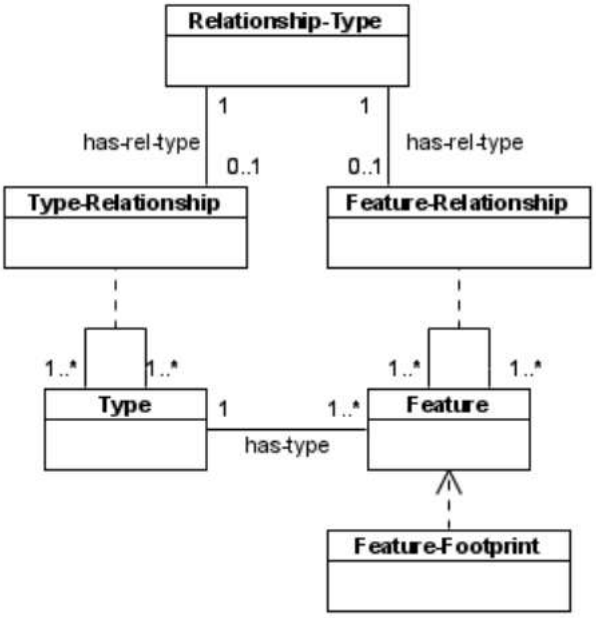

Figura 2.4: Meta-Modelo do GKB 2.0

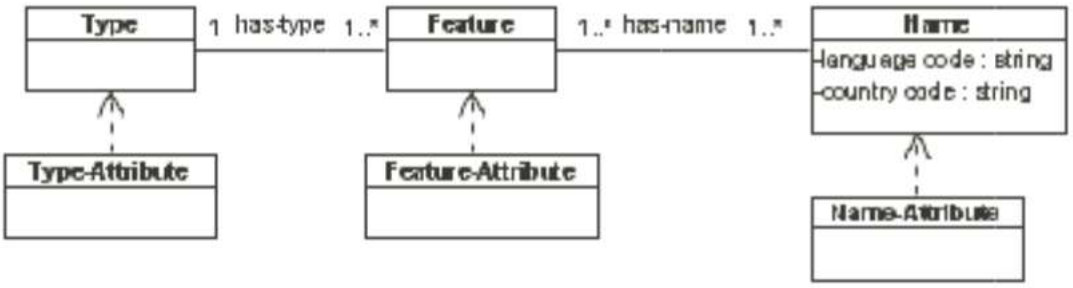

Figura 2.5: Atributos para *Features* e *Types*

ceitos e as *features*, tais como *part-of*, que indica que uma *feature* é parte de outra, ou *adjacent* que indica que duas *features* são adjacentes. Associadas a cada *feature* podem estar *Feature-Footprints*, coordenadas geográficas que representam centróides, caixas delimitadoras ou polígonos.

A Figura 2.5 mostra o modelo anterior estendido com suporte para associar atributos a nomes de entidades geográficas. Os tipos podem ter atributos diferentes, por exemplo, um município tem uma população, uma montanha uma altitude. Uma *feature* tem um nome associado, os nomes poderam ter nomes alternativos, como nomes históricos. As *features* e o seus nomes são classes distintas, cada nome está associado a um ou mais tipos, isto permite a a criação relações 1 para n entre nomes e conceitos geográficos únicos.

Além de modelar as relações entre *features* no mesmo domínio, é possível estabelecer relações entre domínios diferentes. A Figura 2.6 mostra como as relações inter-domínios são representadas. A classe *Adm-Feature* contém os dados administrativos,

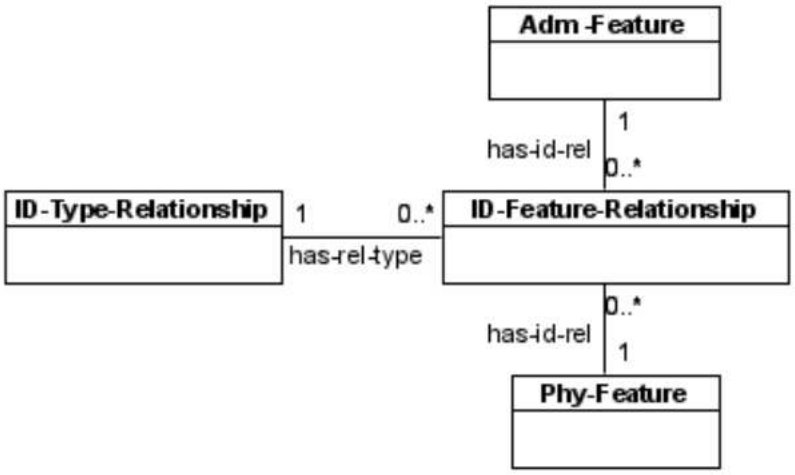

Figura 2.6: Relações entre domínios

Phy-Feature informação do domínio físico. As relações entre ambos são guardadas em *ID-Type-Relationship*, como *part-of* e *adjacency*. Por exemplo os municípios de Lisboa e Setúbal, ambos do domínio administrativo, são adjacentes ao rio Tejo (domínio físico). Outras relações tais como, atravessa, toca, intersecta não são guardadas na classe *ID-Type-Relationship*, mas podem ser inferidas a partir da footprint, por exemplo o rio Douro atravessa o município do Porto, e intersecta o Biótipo Alto Douro Internacional.

2.4.2 Geo-Net-PT

É uma ontologia geográfica pública com âmbito no território de Portugal, tem como base de desenho um repositório baseado num modelo para integração de conhecimento geográfico, GKB. É apresentada no formato *Web Ontology Language* (OWL), o qual é uma recomendação internacional do World Wide Web Consortium (W3C). Pode também ser consultada interactivamente através de uma ligação a uma base de dados. Alternativamente pode-se usar o OWL como base de triplos RDF (W3C, 2004), sujeito-predicado-objecto, e fazer consultas usando a linguagem SPARQL (W3C, 2008).

Na versão actual existem três modelos com dados administrativos, geográficos e da web portuguesa. A informação usada para preencher um modelo GKB com dados do território português é proveniente de diferentes fontes de informação. Nomeadamente, Ministério do Ambiente, Instituto Geográfico do Exército, Instituto Geográfico Português, Instituto da Água, Instituto Nacional de Estatística , Correios e Telégrafos de Portugal e Instituto de Pesquisa da Marinha. A Geo-Net-PT encontra-se disponível por pedido através do endereço http://xldb.di.fc.ul.ptwikiGeo-Net-PT_02. A Tabela 2.7 apresenta a caracterização estatística dos dados administrativos da Geo-Net-PT.

Contém também dados geográficos físicos de Portugal, como rios, serras, albufeiras, parques naturais entre outros dados. Existem outros dados como linhas férreas, hotéis, castelos. A Tabela 2.8 apresenta uma descrição pormenorizada desses dados.

Os diferentes tipos de entidades geográficas presentes na ontologia têm relações entre si, o grafo na Figura 2.7 apresenta o tipo de relações entre possíveis entre conceitos do domínio administrativo, o grafo na Figura 2.8 as relações entre os conceitos do domínio físico. Estas relações são exploradas no processo de desambiguação de significados geográficos dos locais identificados nos textos.

Os conceitos dos dois domínios estão relacionados através do modelo já apresentado na Figura 2.6. Neste momento existem apenas relações do tipo *part-of*, totalizando cerca de 2.752 relações entre conceitos do domínio físico e o domínio administrativo.

2.4.3 WGO - World Geographic Ontology

Está também disponível uma ontologia geográfica com âmbito mundial sob o modelo GKB. A informação foi recolhida de diferentes fontes de informação disponíveis da Web.

Componente	Valor
Conceitos geográficos distintos	388 049
Conceitos geográficos distintos sem CP	199 053
Nomes	265 044
Tipos de conceitos	62
Número de relações	423 836
Número de relações parte-de	386 431
Número de relações de adjacência	33 051
Conceitos do tipo NUT1	3
Conceitos do tipo NUT2	7
Conceitos do tipo NUT3	30
Províncias	11
Distritos	18
Ilhas	11
Concelhos	308
Freguesias	4 260
Zonas	3 594
Localidades	44 386
Arruamentos	146.422
Códigos Postais	187 014
Número de conceitos com dados demográficos (apenas concelhos)	308
Número de conceitos com coordenadas geográficas	4.597
Distritos com coordenadas geográficas	18
Freguesias com coordenadas geográficas	4.260
Concelhos com coordenadas geográficas	308
Ilhas com coordenadas geográficas	11

Tabela 2.7: Caracterização Estatística dos Dados Administrativos

Componente	Valor
Conceitos geográficos distintos	5.662
Nomes	8.250
Tipos de conceitos	21
Número de relações	2.794
Número de relações parte-de	390
Número de relações de adjacência	2.404
Albufeira	90
Aldeia histórica	217
Área protegida	31
Biótopo	58
Castelo	256
Estuário	8
Hotel	381
Linha férrea	38
Marina	26
Monumento natural	5
Museu	507
Nascente	220
Oceano	5
Parque nacional	1
Parque natural	12
Praia	558
Recurso turístico	84
Região natural	305
Rio	2.421
Serra	25
Sítio arqueológico	414
Número de conceitos com coordenadas geográficas	3.208

Tabela 2.8: Caracterização Estatística dos Dados Físicos

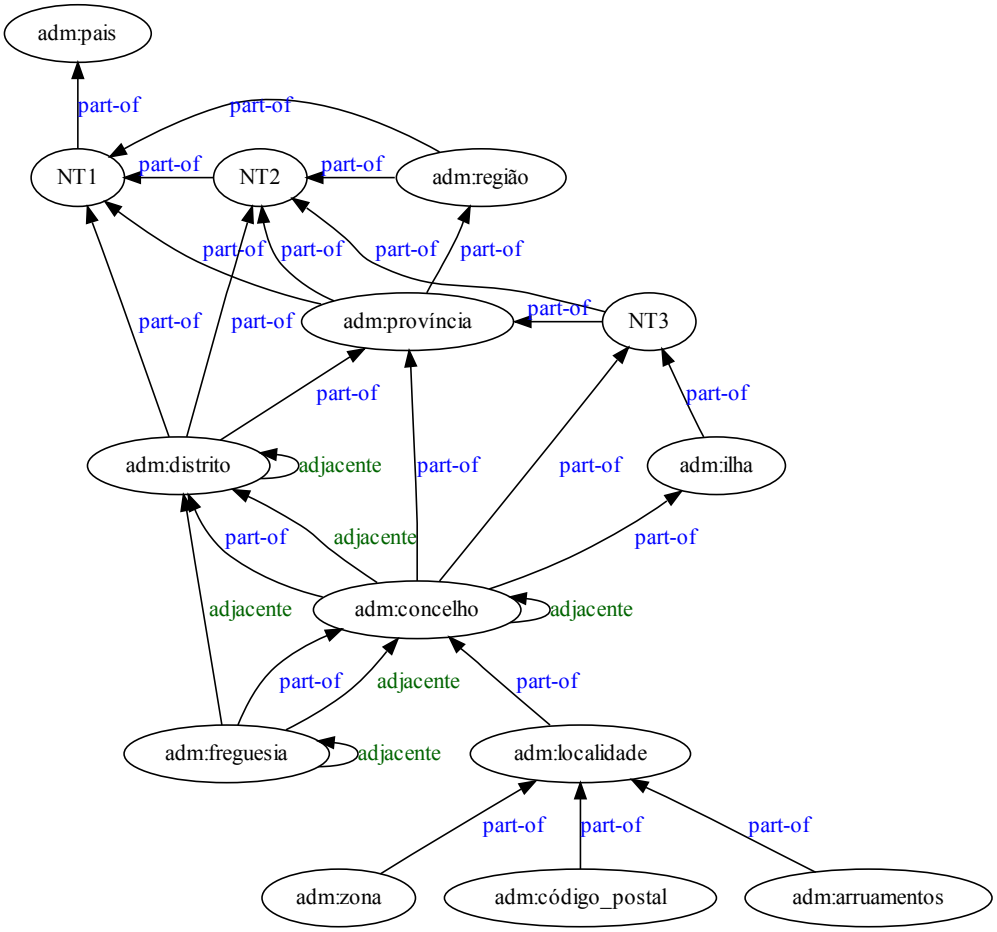

Figura 2.7: Relações entre tipos de conceitos para os dados administrativos

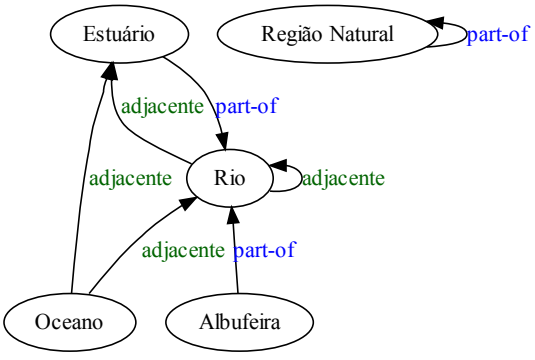

Figura 2.8: Relações entre tipos de conceitos para os dados físicos

As Tabelas 2.9 e 2.9 mostram a caracterização estatística da WGO. Dado que o âmbito geográfico é mundial, o nível de detalhe é muito menor.

O processo de construção, incluindo a limpeza e integração de dados, que deram origem as duas ontologias é descrito por Chaves (2009)

2.4.4 Wiki WGO 2009

Uma terceira ontologia de âmbito mundial foi construida, a Wiki WGO 2009. Tem como base a Wikipedia portuguesa e está organizada segundo uma versão melhorada do GKB (3.0), esta versão do GKB, inclui a geração de ontologias usando dados interligados, e alternativas à utilização de *features* e *feature types* para descrever os recursos (Cardoso et al., 2009).

A construção desta ontologia e o desenvolvimento da nova versão do GKB tiveram como propósito a sua utilização no GikiCLEF 2009 (Santos and Cabral, 2009), um evento de avaliação no âmbito do CLEF (http://www.uni-hildesheim.de/geoclef/) , um evento para avaliação de sistemas de recuperação de informação geográfica. O objectivo é avaliar sistemas que encontram documentos ou artigos na Wikipedia que contêm resposta a uma determinada pergunta ou informação. O processo de procura da resposta envolve de alguma forma raciocínio geográfico por parte dos sistemas.

2.5 Similaridade Semântica

O processo de desambiguação de entidades geográficas poderá ser feito também recorrendo a medidas de similaridade semântica. Uma outra tarefa do projecto GREASE foi o

Componente	Valor
Conceitos geográficos distintos	12.982
Nomes	12.102
Tipos de conceitos	7
Número de relações	23.732
Número de relações parte-de	12.562
Número de relações de adjacência	11.170
Conceitos do tipo ISO-3166-1	486
Conceitos do tipo ISO-3166-2	3.977
Cidades Capitais	464
Lugares	4024
Divisões Administrativas	3.216
Aglomerações Populacionais	752
Regiões	63

Tabela 2.9: Caracterização Estatística dos dados administrativos na WGO

Componente	Valor
Conceitos geográficos distintos	721
Nomes	750
Tipos de conceitos	17
Número de relações	525
Número de relações parte-de	513
Número de relações de adjacência	12
Continentes	16
Mares	8
Lagos	67
Oceanos	3
Ilhas	215
Rios	88
Estreitos	2
Cordilheiras	4
Montanhas	85
Circuitos	98
Aeroportos	113
Catedrais	3
Canais	2
Desertos	1
Penínsulas	2
Túneis	12

Tabela 2.10: Caracterização Estatística dos dados físicos na WGO

estudo de medidas de similaridade semântica na Geo-Net-PT.

A Similaridade Semântica é usada para determinar quão semelhantes são dois conceitos dentro de um domínio, com base nas suas propriedades semânticas. A Geo-Net-PT pode ser representada como um grafo directo acíclico, estruturado em forma de árvore, ou seja é possível para todos os conceitos, com excepção dos termos folha, saber quem são os seus descendentes. Os nomes dos conceitos na Geo-Net-PT foram anotados com a frequência da sua ocorrência num dado corpus que serviu para calcular o *Information Content* (IC), descrito a seguir:

2.5.1 Information Content

O IC é um atributo numérico dado a cada conceito administrativo contido na Geo-Net-PT. É definido com base na frequência do seu nome e frequência dos nomes dos seus descendentes num mesmo corpus:

$$HFreq(c) = Freq(c) + Freq(Descendentes(c))$$

$Prob(c)$ define a probabilidade de haver uma referência a um conceito geográfico num texto, com base nos seus descendentes, e nas frequências calculadas. Mesmo que um conceito não ocorra explicitamente num dado texto, existe sempre uma probabilidade associada se pelo menos o dos seus descendentes ocorrer.

$$Prob(c) = \frac{HFreq(c)}{maxFreq}$$

$maxFreq$ é a frequência máxima de todos os conceitos definidos na Geo-Net-PT, ou seja é a frequência do nó raiz. Quanto mais descendentes um conceito tiver menos informação expressa, conceitos que são folhas no grafo representado pela Geo-Net-PT são mais específicos geograficamente, sendo que a informação que estes expressão é máxima, assim IC é definido como:

$$IC(c) = -logProb(c)$$

2.5.2 Medidas de Similaridade Semântica

O IC dos conceitos do domínio administrativo é utilizado para calcular a similaridade semântica entre dois conceitos geográficos. Uma função de medida de similaridade semântica, recebe o IC de dois conceitos e devolve um valor real entre 0 e 1. Quanto mais próximo de 1 mais alta a similaridade entre os dois conceitos o que significa que os dois conceitos estão geograficamente relacionados.

A similaridade semântica é uma alternativa às técnicas de desambiguação, quando se tem que seleccionar uma de entre todas os referências ontológicas correspondentes a entidades extraídas de um texto. Por exemplo, tendo sido extraídos os termos ”Lisboa” e ”Santa Catarina”, estes podem ter os seguintes correspondentes na Geo-Net-PT – na realidade são muito mais, mas por simplicidade apenas se apresentam estes:

- Lisboa como Concelho (#146)
- Lisboa como Localidade no Concelho de Monção (#379800)
- Santa Catarina como Freguesia no Concelho de Lisboa (#418458)
- Santa Catarina como Rua no Concelho Porto (#295404)

Aplicando uma função de medida de semelhança semântica: $SSM(IC_1, IC_2) \in [0, 1]$ em cada par, escolhe-se o par com o IC mais alto.

$$SSM\ (146, 418458) = 0.5849326208368193$$
$$SSM\ (146, 295404) = 0.06534881335785453$$
$$SSM\ (379800, 418458) = 0.06376224760427719$$
$$SSM\ (379800, 295404) = 0.1414917751967333$$

Neste exemplo, o par ”Lisboa, Concelho” (146) e ”Santa Catarina, Freguesia” (418458) tem o valor mais alto, significando que geograficamente são o par mais relacionado, descarta-se assim as outras referências ontológicas correspondentes a ”Lisboa” e ”Santa Catarina”.

2.6 Sumário

Nesta secção foram apresentados os recursos e as tecnologias utilizados para desenvolver o HENDRIX. A extracção de entidades tem por base os *Conditional Random Fields*, e a Colecções Dourada do HAREM foi o recurso utilizado para fazer o treino do modelo gerado. A validação e desambiguação das entidades extraídas é conseguida com recurso a duas ontologias geográficas, e aplicando heurísticas de desambiguação já utilizadas em trabalhos anteriores. Uma segunda alternativa às heurísticas de desambiguação, são as medidas de similaridade semântica aplicadas à Geo-Net-PT. Na próxima secção o sistema HENDRIX é descrito em detalhe.

Capítulo 3

HENDRIX

Este capítulo descreve o HENDRIX, o sistema desenvolvido, nos vários módulos que o constituem e a sua arquitectura. É descrito o processo de transformação das Colecções Douradas (CD) de forma a que estas contenham apenas entidades anotadas pertencentes à categoria LOCAL, para depois seram usadas na geração do modelo *Conditional Random Fields* (CRF). São descritas as funcionalidades do módulo de software desenvolvido por mim, que integra o sistema HENDRIX, o PAREDES. No final da secção é apresentado o PAGE, responsável por extracção de entidades para colecções de documentos.

3.1 Arquitectura

O HENDRIX (acrónimo de **H**endrix is an **E**ntity **N**ame **D**esambiguator and **R**ecognizer for **I**nformation **E**xtraction), é o sistema que desenvolvi para extrair entidades geográficas de documentos em português e produzir o seu resumo geográfico. É constituído por:

- um módulo de aprendizagem supervisionada de sequências de termos em texto denotando nomes de locais, baseado no modelo de *Conditional Random Fields* (CRF) implementado pelo Minorthird (Cohen, 2004);
- um módulo de software, PAREDES, desenvolvido para a análise e referenciação dos nomes das entidades extraídas a referências geográficas.

A Figura 3.1 apresenta a arquitectura do sistema. O HENDRIX recebe de entrada um documento, já pré-processado, contendo apenas com texto, sem meta-dados extra ou etiquetas HTML ou XML. O documento é passado ao Minorthird para efectuar a extracção. Este analisa o documento e devolve um ficheiro com as entidades extraídas, e posições no texto onde ocorrem. A extracção é feita pelo Minorthird recorrendo a um modelo de CRF treinado a partir de documentos com esses nomes de entidades anotados.

Esta informação, os nomes de entidades extraídas e as suas posições no texto são depois passados ao PAREDES, que inicia o processo de validação, usando as ontologias

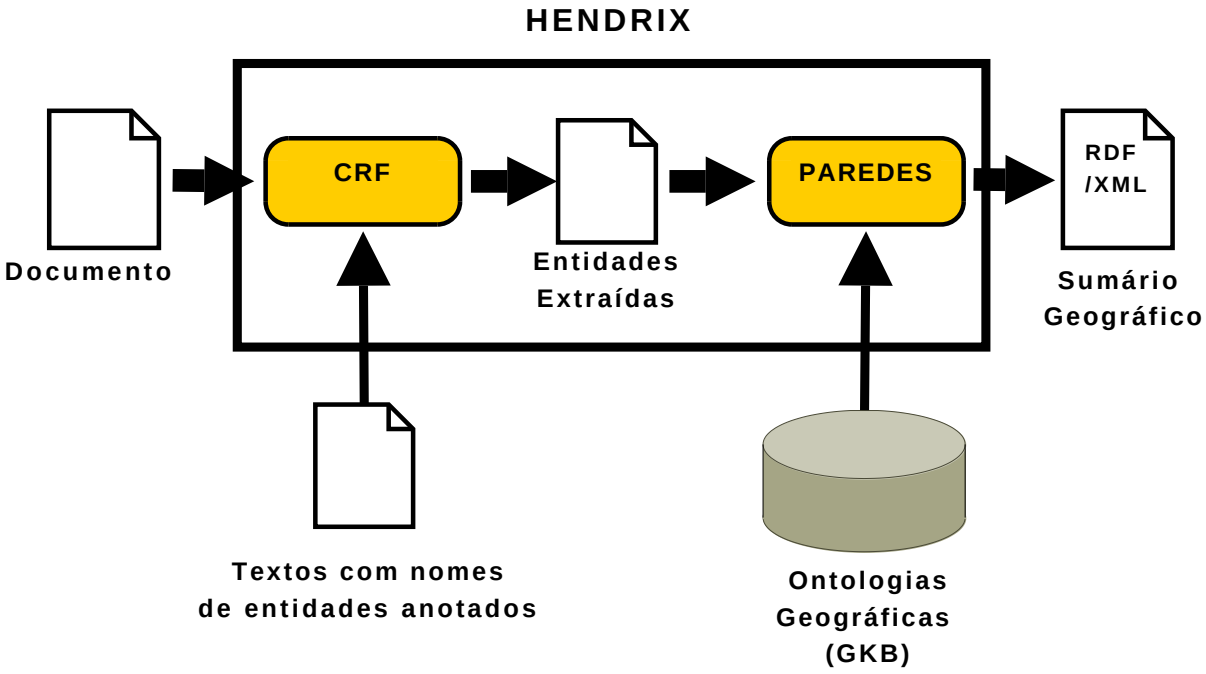

Figura 3.1: Arquitectura geral do sistema HENDRIX

geográficas, de forma a comprovar a informação extraída pelo modelo de CRF gerado. Depois de validadas as entidades extraídas, é iniciado o processo de desambiguação. Das entidades que encontradas na Geo-Net-PT, são extraídos todos os significados geográficos possíveis, e destes é necessário seleccionar os que de facto são referidos no texto.

Existe ainda um outro módulo, PAGE, construído com base no Hadoop (Dean and Ghemawat, 2004), que permite aplicar o modelo de CRF em larga escala, recorrendo a um *cluster* de computadores, permitindo assim fazer a extracção de entidades para colecções de documentos (ver Figura 3.2).

3.2 Geração do modelo CRF

As CD do HAREM foram o artefacto usado para gerar o modelo baseado em CRF para extrair nomes de entidades geográficas. No HAREM as entidades da categoria ”LOCAL” podem estar classificadas em três tipos: ”FISICO”, ”HUMANO” ou ”VIRTUAL”. As entidades com tipo ”VIRTUAL” dizem respeito a sítios abstractos com função de alojamento de conteúdos, tais como jornais, endereços electrónicos, ou programas de televisão. Não correspondem a qualquer localização física, e uma vez que estes tipo de locais não têm interesse no contexto de entidades geográficas, estas entidades não fazem parte da CD alterada que serviu para treinar o CRF usado para detecção de nomes geográficos.

Atendendo a que as CD são etiquetadas em XML, foi desenvolvido um *script* em XSLT para fazer a transformação das CD originais de forma a que estas ficassem eti-

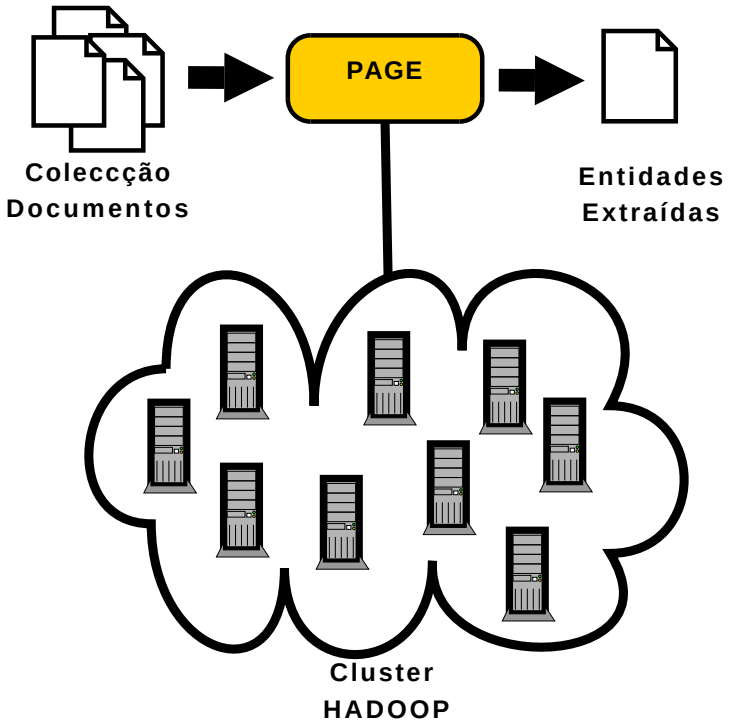

Figura 3.2: Arquitectura do módulo PAGE

quetadas apenas com entidades da categoria LOCAL, pertencentes ao tipo HUMANO ou FISICO, deixando todas as outras sem qualquer tipo de anotação.

A Tabela 3.1 mostra a caracterização das CD para entidades da categoria ”LOCAL” no que respeita ao número e tamanho das colecções HAREM.

	MiniHAREM	HAREM I	HAREM II
Tamanho	514 Kbytes	734 Kbytes	1.1 Mbytes
Nº Entidades Únicas	397	514	612
Total	792	1146	1200

Tabela 3.1: Caracterização das CD para a categoria LOCAL

3.3 PAREDES

O PAREDES (acrónimo de **P**AREDES **A**dvocates **R**ecognized **E**ntities for **D**esambiguation and **E**xtraction of **S**ummaries) é o módulo de *software* desenvolvido para classificar as entidades extraídas pelo Minorthird. As suas tarefas principais são:

- emparelhamento das entidades extraídas com conceitos nas ontologias geográficas suportadas pelo HENDRIX;

n_name	n_ascii_name	n_cap_name
alcácer do sal	alcacer do sal	Alcácer do Sal
dão-lafões	dao-lafoes	Dão-Lafões
são joão de negrilhos	sao joao de negrilhos	São João de Negrilhos

Tabela 3.2: Exemplo de representações de nomes alternativos nas ontologias do sistema GKB.

- desambiguação dos possíveis significados para as entidades extraídas;
- geração de resumos geográficos;

3.3.1 Processo de Emparelhamento

No PAREDES, os nomes de entidades geográficas são utilizados em consultas feitas às ontologias geográficas usadas para georeferenciação. Dado que os textos de onde são extraídas são provenientes da web, a presença de erros ortográficos, a falta de acentuação, ou a não capitalização de locais pode ocorrer com frequência. Antes de serem utilizadas nas interrogações as entidades são transformadas.

Para fazer essa identificação, a cadeia de caracteres correspondente a cada entidade extraída é toda convertida em caracteres minúsculos. Nas ontologias usadas pelo PAREDES, cada nome presente nas ontologias é representado de três formas diferentes: apenas em letras minúsculas e com acentuação; em minúsculas e sem acentuação; e com os nomes capitalizados e com acentuação. A Tabela 3.2 mostra um exemplo da representação dos nomes.

Ao transformar as entidades todas para minúsculas e procurando um correspondente nas representações dos nomes em minúsculas com ou sem acentuação evitam-se as falhas devido a diferenças por maiusculização, ou falta de acentuação. No entanto o problema com erros ortográficos não fica resolvido, é necessário para isso uma abordagem diferente da simples comparação de caracteres, por exemplo, por distância de edição (Levenshtein, 1966).

Além de conversão das entidades para caracteres minúsculos, são aplicadas expressões regulares para fazer a detecção de abreviaturas, que são expandidas depois de detectadas. A expansão permite completar as abreviaturas encontradas em algumas entidades extraídas. Presentemente são expandidas as abreviaturas apresentadas na Tabela 3.3. Isto é necessário para encontrar conceitos geográficos correspondentes às entidades extraídas com abreviaturas, pois os nomes em ontologias como a Geo-Net-PT são guardados sem abreviaturas.

As consultas feitas à ontologia geográfica são de dois tipos. Primeiro a interrogação é feita com a entidade extraída, apenas convertida para caracteres minúsculos e com as abreviaturas expandidas. Depois é aplicada uma expressão regular para detectar a presença

Abreviatura	Entidade extraída	Abreviatura expandida
D.	Rua D. Afonso Henriques	Rua Dom Afonso Henriques
S.	S. Pedro de Moel	São Pedro de Moel
St^{o}	Vila Real St^{o} António	Vila Real Santo António
St^{a}	St^{a} Engrácia	Santa Engrácia
Sta.	Sta. Eufémia	Santa Eufémia
Sto.	Sto. António dos Cavaleiros	Santo António dos Cavaleiros

Tabela 3.3: Exemplos de abreviaturas expandidas.

```
^(acesso|adro|alameda|arruamento|avenida|azinhaga|bairro|
beco|cais|calçada|caminho|campo|canada|canto|carreira|concelho|
código postal|distrito|entidade geográfica administrativa|
escadas|escadinhas|estrada|freguesia|ilha|jardim|ladeira|
largo|localidade|loteamento|lugar|monte|nut|nut1|nut2|nut3|
outro|pais|parque|passeio|ponte|praceta|praça|província|
pátio|quelha|quinta|rampa|recanto|região|rotunda|rua|ruela|
sítio|terreiro|travessa|urbanização|vale|vereda|via|viela|
zona|av\.|av|av[a]|av[a]\.|pra\.|pra)(\s|\sd[aeo]s).*+
```

Figura 3.3: Expressão regular utilizada para detectar tipos de conceitos

de um tipo de conceito na entidade extraída. O objectivo é conseguir separar no nome de entidade extraída o tipo de conceito e o nome da entidade. Por exemplo, na entidade extraída ”Avenida da Liberdade” aplicando uma expressão regular detecta-se o conceito ”Avenida”. A Figura 3.3 apresenta a expressão regular usada quando o processamento é feito com a Geo-Net-PT. Esta é carregada com os *feature types* presentes na ontologia, e mais algumas abreviaturas de tipos de conceitos, adicionadas manualmente. Este tipo de abreviaturas ocorre com alguma frequência nas entidades extraídas.

Os nomes de entidades geográficas e os tipos de conceitos a que podem corresponder são duas classes diferentes no modelo GKB 2.1, podendo assim o mesmo nome representar diferentes tipos de conceitos. Por exemplo, o nome ”Liberdade” pode representar até 486 conceitos geográficos diferentes, como ruas, travessas, avenidas, largos, etc.

De seguida são aplicadas outras expressões regulares que separam o tipo de conceito e o nome da entidade geográfica. Artigos definidos compostos situados entre o tipo de conceito e nome – *da, de, do, das, dos* – quando presentes são também retirados, dado que estes não fazem parte da representação dos nomes na Geo-Net-PT. A Tabela 3.4 mostra alguns exemplos que resultam da extracção de tipos de conceitos geográfico e nomes de uma entidade extraída.

Tendo extraído o nome e o tipo de conceito é possível interrogar a ontologia com mais especificidade. Por exemplo, em ”Avenida da Liberdade”, pedindo todos as re-

Nome de Entidade Extraída	Tipo de CG	Nome de Entidade (na ontologia)
Avenida António Augusto de Aguiar	Avenida	António Augusto de Aguiar
Avenida de Roma	Avenida	Roma
Av. do Brasil	Avenida	Brasil
Av. Fernão de Magalhães	Avenida	Fernão de Magalhães
Av.ª dos Aliados	Avenida	Aliados
Av.ª 5 de Outubro	Avenida	5 de Outubro
Av.Calouste Gulbenkian	Avenida	Calouste Gulbenkian
Pra 25 de Março	Praça	25 de Março
Distrito de Lisboa	Distrito	Lisboa
Rua de S. Bento	Rua	São Bento
Largo da Misericórdia	Largo	Misericórdia

Tabela 3.4: Separação entre o tipo de conceito e o seu nome

ferências cujo o nome é ”Liberdade” e que têm como tipo de conceito ”Avenida”. Este tipo de interrogações, com um tipo de conceito associado, permite reduzir o número de referências potencialmente emparelháveis, facilitando o processo de desambiguação.

Como exemplo, as referências geográficas com o nome ”Liberdade” na Geo-Net-PT são 486. Se a entidade tiver o conceito ”Avenida”, reformulando a consulta, de forma a pedir todas as referências com nome ”Liberdade” e tipo de conceito ”Avenida” o conjunto de resultados é reduzido para 69, representando todos os conceitos geográficos que se referem a avenidas em Portugal com o nome ”Liberdade”.

Sempre que é detectado um tipo de conceito, uma nova consulta é feita, usando o tipo de conceito como parte do nome , independentemente de a consulta anterior com o mesmo nome de entidade extraída ter devolvido referencias ou não. Um dos problemas existentes são referências geográficas que têm no nome tipos de conceitos geográficos, por exemplo:

”Avenida 24 de Julho”: ao ser feita uma primeira consulta sem usar o tipo de conceito ”Avenida” é devolvida uma referência que tem como tipo de conceito geográfico Zona. No entanto o mais provável é a entidade extraída estar a referir-se a uma referência a um referência com tipo de conceito Avenida e com o nome ”24 de Julho”.

”Ponte de Lima”: ao fazer primeiro a consulta utilizando o tipo de conceito associado, ”Ponte”, não é devolvida nenhuma referência, no entanto a entidade poderá ser uma referência ao concelho de ”Ponte de Lima”.

Desta forma, são sempre feitas duas consultas para as entidades com um tipo de conceito geográfico associado, uma usando-o no nome, outra usando-o como tipo de conceito

geográfico. Para cada entidade extraída de um documento, são guardadas as posições no documento onde esta ocorre e os identificadores correspondentes na ontologia usada para georeferênciação.

Cache

Para cada entidade processada em cada documento e com pelo menos um correspondente nas ontologias são guardadas as suas referências numa *cache* de entidades resolvidas. Uma outra *cache* de entidades não resolvidas é usada para entidades não encontradas em nenhuma das ontologias. Isto permite que, sempre que haja uma entidade a ser processada que seja repetida, evitar uma nova consulta às ontologias, o que acelera bastante o processo. Basta consultar as *caches*, e extrairem-se as referências caso esteja nas entidades resolvidas, ou marcar-se como não tendo um correspondente, caso esteja na *cache* de entidades não resolvidas.

Todas as referências encontradas na Geo-Net-PT partir das entidades extraídas são guardadas para serem usadas no processo de desambiguação. A WGO e a Wiki WGO 2009 apenas foram usadas como um dicionário de nomes extra. As entidades que não são encontradas na Geo-Net-PT mas estão na WGO ou na Wiki WGO 2009 contam como entidades resolvidas, mas não fazem parte do processo de desambiguação.

3.3.2 Processo de Desambiguação

No processo de desambiguação tenta-se seleccionar de todas as referências geográficas encontradas na Geo-Net-PT aquela a que o documento realmente se refere. É um processo de filtragem, numa primeira fase reduz-se o número de referências, numa segunda fase através da exploração de relações ou usando medidas de semelhança semântica, tenta-se desambiguar as possíveis referências para uma dada entidade. Os métodos utilizados são descritos a seguir.

Redução de referências geográficas

Uma entidade geográfica extraída de um texto e com representação na ontologia poderá estar associada a várias referências geográficas. Isto acontece com frequência quando uma entidade é extraída sem nenhum tipo de conceito associado. Há casos em que são extraídos nomes de países, e onde as interrogações feitas à Geo-Net-PT devolvem conceitos do tipo arruamento – avenidas, largos, praças, ruas, alamedas, etc – por exemplo, fazendo uma consulta por ”Brasil” são devolvidas 83 referências a arruamentos. O mesmo acontece com o nome de subdivisões internacionais, por exemplo: ”Londres” ou ”Madrid” correspondem cada a 8 conceitos geográficos do tipo arruamento, e também com subdivisões nacionais. ”Beja” poderá corresponder a 1 Distrito, 1 Concelho, 3 Freguesias ou 15 arruamentos.

De forma a reduzir o número de entidades geográficas emparelhadas, várias heurísticas são aplicadas, conforme as referências extraídas:

- Quando só se está a gerar resumos geográficos com âmbito no território português, só são tidas em consideração apenas entidades que têm referências na Geo-Net-PT.
- Aplica-se a heurística de um referente por documento. Se a mesma entidade é referida várias vezes no documento, assume-se que é sempre referenciada à mesma entidade geográfica.
- Para as entidades extraídas com um tipo de conceito associado, todas as referências encontradas na ontologia são utilizadas. Extraindo ”Travessa de Tomar” e identificando o tipo de entidade geográfica, ”Travessa”, usam-se todas os identificadores.
- Se uma entidade foi extraída sem nenhum tipo de conceito associado e as referências encontradas incluem subdivisões e arruamentos escolhem-se apenas as subdivisões, eliminando as referências mais baixas na hierárquica, como os arruamentos. Extraindo ”Tomar” usam-se apenas os identificadores da ontologia que correspondem a subdivisões, excluindo os identificadores que representam ”Travessa de Tomar” ou outro tipo de arruamentos cujo o nome é ”Tomar”.
- Se uma entidade foi extraída sem nenhum tipo de conceito associado e apenas há referências a arruamentos, as referências são descartadas. Extraindo ”Brasil” são devolvidas 83 referências a arruamentos, neste caso nenhuma é utilizada.

As heurísticas partem do princípio que quando há uma referência a um arruamento num texto, este é feito explicitamente, ou seja se num texto houver uma referência à ”Praça de Londres”, o tipo de conceito, neste caso ”Praça”, é referido.

Identificação de relações

As referências geográficas presentes nas ontologias estão agrupadas em conceitos, e estes têm relações definidas entre si. As relações procuradas entre as referências geográficas que representam as entidades extraídas seguem a estrutura do grafo de relações apresentado na Figura 2.7. Na Geo-Net-PT as relações extraídas são as seguintes:

- A relação *adjacente-a* existe entre dois conceitos, e designa que a área geográfica de dois conceitos são vizinhas.
- A relação *parte-de* existe quando um conceito geográfico está contido noutro.
- A relação *filho-de* extrai-se através da transitividade da relação *parte-de*, por exemplo:

  Se (**A** *parte-de* **B**) e (**B** *parte-de* **C**) então (A *filho-de* **C**);

Apenas as relações *adjacente-a* e *filho-de* são exploradas, dado que a relação *parte-de* é um caso particular da *filho-de*. São exploradas relações entre os seguintes tipos de conceitos:

- (Freguesia) *adjacente-a* (Freguesia)
- (Freguesia) *adjacente-a* (Concelho)
- (Freguesia) *adjacente-a* (Distrito)
- (Concelho) *adjacente-a* (Concelho)
- (Concelho) *adjacente-a* (Distrito)
- (Distrito) *adjacente-a* (Distrito)
- (Arruamentos,Zona) *filho-de* (Local,Concelho,Distrito,Ilha,Província,Região,NT3,NT2,NT1)
- (Freguesia,Local) *filho-de* (Concelho,Distrito,Ilha,Província,Região,NT3,NT2,NT1)
- (Concelho) *filho-de* (Distrito,Ilha,Província,Região,NT3,NT2,NT1)
- (Distrito,Ilha) *filho-de* (Província,Região,NT3,NT2,NT1)
- (NT3) *filho-de* (Província,Região,NT3,NT2,NT1)
- (Província) *filho-de* (Região,NT2,NT1)
- (Região) *filho-de* (NT2,NT1)
- (NT2) *filho-de* (NT1)

Para cada relação encontrada é guardado o tipo de relação e a distância. No caso da relação de adjacência a distância é 1.

Heurísticas de Desambiguação e Inferência de Âmbito Geográfico

Foram desenvolvidas 3 heurítiscas de desambiguação para inferência do âmbito geográfico e geração de resumos geográficos. Estas heurísticas têm com objectivo de entre todas as referências geográficas da ontologia que as entidades geográficas extraídas representam, seleccionar aquelas que de facto são as referidas no texto. Essas referências seleccionadas são usadas para calcular o âmbito geográfico do documento.

1. São extraídas todas as relações possíveis entre as referências encontradas, eliminando as que não têm relações. É escolhida a referência que mais relações com outras referências agrega, ou seja a que mais relações tem. Se houver mais do que uma referência com o mesmo número de máximo relações, procura-se na ontologia o

antecessor comum a elas. A referência escolhida, ou o antecessor comum no caso de mais do que uma, definem o âmbito geográfico do documento.

2. As medidas de semelhança são usadas para desambiguar às entidades extraídas de um documento. São aplicadas às entidades pela ordem de ocorrência no documento. De seguida determina-se o antecessor comum na ontologia mais próximo das entidades desambiguadas. O antecessor comum define o âmbito geográfico do documento.

   Por exemplo, tendo o seguinte texto: "...deslocou-se pela Avenida da República em direcção ao Marquês de Pombal, aí apanhou o metro em direcção ao Rossio". Extraindo as entidades: "Avenida da República", "Marquês de Pombal" e "Rossio". Calcula-se a medida de semelhança entre os vários identificadores para "Avenida da República" e "Marquês de Pombal", escolhendo-se os dois com o valor mais elevado. De seguida calcula-se a medida de semelhança entre o identificador escolhido para "Avenida da República" e os vários identificadores para "Marquês de Pombal". Ao final de entre os três identificadores procura-se o antecessor comum dos três mais próximo. O antecessor comum aos três identificados define o âmbito geográfico do documento.

3. É semelhante à heurística anterior, com a diferença de que em vez de se calcular o antecessor comum de entre as referências desambiguadas para inferir o âmbito geográfico, extraem-se as relações entre as entidades desambiguadas. A referência que mais relações, ou antecessor comum, no caso de haver mais do que uma com o número máximo de relações, definem o âmbito geográfico do documento.

A pesquisa ao final, pelo antecessor comum mais próximo, permite associar o documento a locais que não apareçam explicitamente descritos no texto, por exemplo, no exemplo da segunda heurística, poderia-se chegar à referência na ontologia de Lisboa como Concelho, sem que o nome "Lisboa" apareça explicitamente no texto.

3.3.3 Geração de Resumos Geográficos

O principal objectivo da extracção das entidades geográficas dos documentos é a sua utilização para outras aplicações, como recuperação de informação ou visualização de informação geo-referenciada. Os resumos geográficos apresentam as entidades extraídas e desambiguadas tendo em conta que seram utilizados por outras aplicações.

O resumo geográfico descreve as entidades extraídas do texto, indicando o número de ocorrências, e as referências geográficas que correspondem na ontologia. Para cada entidade extraída e identificada na Geo-Net-PT, existe apenas uma única referência na Geo-Net-PT. São também indicadas as entidades geográficas extraídas, mas que foram eliminadas no processo de desambiguação. São também apresentadas as entidades extraídas mas não resolvidas na Geo-Net-PT

Os resumos geográficos são apresentados com base no formato de triplos Resource Description Format (RDF), apresentado a semântica geográfica do documento de forma ser processada.

3.4 Processamento de Colecções de Documentos

Nesta secção é descrito o processo usado para fazer a extracção de entidades geográficas para uma grande colecção de documentos. Mais concretamente para a WPT05, uma recolha da Web portuguesa, totalizando cerca de 40 Gigabytes de dados. A extracção foi feita com recurso ao Hadoop, uma *framework* para processamento distribuído.

3.4.1 PAGE

O **PAGE** (acrónimo para **P**age **A**cquires **G**eographic **E**ntities) foi desenvolvido sobre o Hadoop de forma a poder aplicar o modelo de CRF treinado para fazer extracção de entidades geográficas a uma recolha da Web portuguesa. O desenvolvimento do PAGE teve em consideração o formato dos dados usado para descrever os documentos que fazem parte da WPT05 e o paradigma de MapReduce do Hadoop. Segue-se uma descrição do formato RDF usado na WPT05 e do paradigma MapReduce.

Formato dos documentos da WPT05

A recolha da Web portuguesa encontra-se no formato RDF. O formato dos ficheiros RDF encontra-se exemplificado na Figura 3.4. Os meta-dados na WPT05 descrevem vários atributos do documento, como o endereço IP do servidor onde foram recolhidos, o servidor HTTP utilizado, entre outros. O texto do documento encontra-se dentro da etiqueta `<wpt:filteredText>`, O atributo `rdf:about` da etiqueta `<rdf:Description>`, guarda o URL do documento. Caso o documento seja um duplicado, a etiqueta é substituída por outra, `<wpt:duplicateOf>` indicando o URL de que o documento é cópia. A língua em que o documento se encontra escrito é assinlado pela etiqueta `<dc:language>`.

Hadoop

O Hadoop é uma plataforma com suporte para processamento distribuído de dados. Permite que uma aplicação seja executada sob várias unidades de processamento de forma a poder lidar com grandes quantidades de dados. Incluí mecanismos para distribuição do processamento por todo o *cluster* de unidades de processamento, um sistema de ficheiros de inspirado no *Google File System* (Ghemawat et al., 2003), a monitorização do processamento poderá ser feita usando um interface Web. Permite o desenvolvimento de aplicações com base no paradigma MapReduce (Dean and Ghemawat, 2004), em que é especificado uma função $map()$ que processa um par $(chave1, valor1)$ gerando um outro

```
<rdf:Description rdf:about="http://egasmoniz.blogspot.com/2005/02/o-
    poltico-na-sombra-do-cientista.html">
    <ore:isAggregatedBy rdf:resource="http://egasmoniz.blogspot.com
        /2005/02"/>
    <wpt:ipAddr rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
     66.102.15.101
    </wpt:ipAddr>
    <wpt:server rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
     apache
    </wpt:server>
    <wpt:statusCode rdf:datatype="http://www.w3.org/2001/XMLSchema#int"
        >
     200
    </wpt:statusCode>
    <dcterm:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#
        dateTime">
     2005-03-14T00:00:00Z
    </dcterm:modified>
    <wpt:fetched rdf:datatype="http://www.w3.org/2001/XMLSchema#
        dateTime">
     2005-07-19T12:39:48Z
    </wpt:fetched>
    <dc:format rdf:resource="text/html"/>
    <wpt:arcName rdf:resource="WPT-9-20080820064526-00126"/>
    <wpt:filteredText>Egas Moniz: O politico na sombra do cientista
        Blogger Get your own blog Next blog BlogThis! Egas Moniz
Blogue destinado a comparar, incluir, discutir, divulgar e criticar
    analises, testemunhos, bibliografias e opinioes acerca de Egas
    Moniz, vida, obra e tudo mais que cada um achar relevante para o
    conhecimento do primeiro Nobel portugues de Medicina ou Fisiologia.
Segunda-feira, Fevereiro 28, 2005 O politico na sombra do cientista....
   </wpt:filteredText>
   <dc:language>pt</dc:language>
</rdf:Description>
```

Figura 3.4: Exemplo de um RDF que descreve um documento

conjunto de pares $(chave2, valor2)$, e uma função $reduce()$ que funde todos os valores intermediários com a mesma chave, conceptualmente:

$$map(k1, v1) -> list(k2, v2)$$
$$reduce(k2, v2) -> list(v2)$$

A Figura 3.5 mostra o fluxo de execução de um programa sobre a plataforma Hadoop.

Funcionamento do PAGE

O Hadoop permite que se defina como os dados de entrada são partidos de modo a serem processados em paralelo pelas várias funções $map()$. Desta forma o PAGE recebe à entrada os vários ficheiros RDF que constituem a WPT05. Analisa o RDF de forma a passar à função $map()$ um documento da WPT05, definido pela etiqueta `<rdf:Description>`.

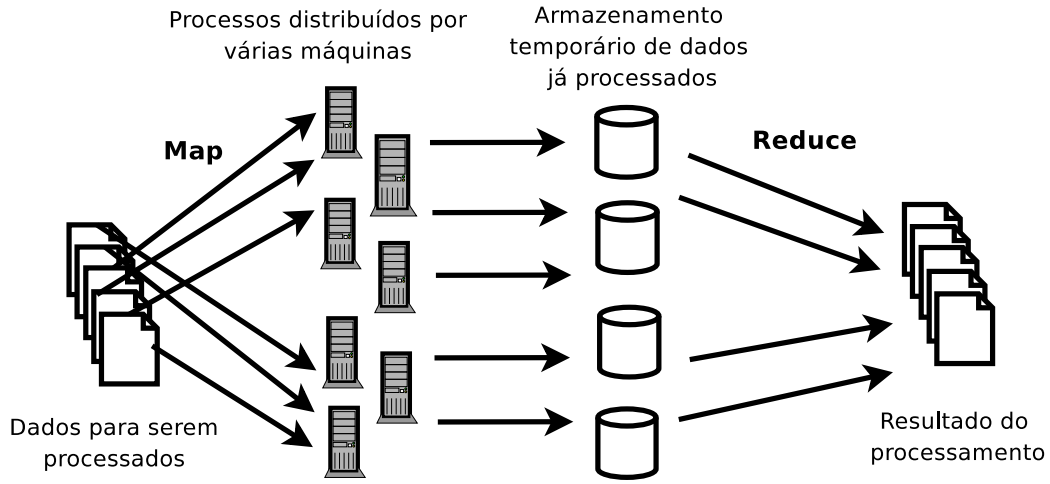

Figura 3.5: Fluxo de processamento de dados na plataforma HADOOP

Apenas são processados documentos contendo a etiqueta `<filteredText>`, pois existem documentos que não contêm texto nenhum, e a etiqueta `<dc:language>` com o valor `pt`, correspondendo a documentos em português. Cada um destes documentos foi processado um objecto Java, o modelo de CRF gerado pelo Minorthird. São criadas várias instâncias da função $map()$, tantas quanto o número de unidades de processamento disponíveis.

Não foi necessário recorrer à fase de $reduce()$, pois a função $map()$ recebe como entrada um documento, e produz como saída as possíveis entidades geográficas identificadas no documento. Estes dados são gravados num ficheiro de saída

Desta forma para cada ficheiro RDF processado é produzido um ficheiro pelo Hadoop com o formato apresentado na Figura 3.6 contendo as entidades geográficas identificadas para cada documento. Cada linha apresenta o URL do documento, a posição no texto onde a entidade começa, a posição onde acaba, e a entidade extraída. Os espaços em branco separam os diferentes documentos que fazem parte do mesmo RDF.

```
http://afilosofia.no.sapo.pt/12Hegel.htm|469|474|Berna
http://afilosofia.no.sapo.pt/12Hegel.htm|517|527|Frankfourt
http://afilosofia.no.sapo.pt/12Hegel.htm|716|722|Berlim

http://afinalvoltei.blogspot.com/|7430|7438|Portugal
http://afinalvoltei.blogspot.com/|7753|7765|Rio Guadiana
http://afinalvoltei.blogspot.com/|7786|7791|Elvas
http://afinalvoltei.blogspot.com/|7885|7893|Guadiana
http://afinalvoltei.blogspot.com/|7931|7938|Espanha
http://afinalvoltei.blogspot.com/|7972|7980|Olivenca
http://afinalvoltei.blogspot.com/|8003|8015|S. Francisco
http://afinalvoltei.blogspot.com/|8028|8037|Vila Real
```

Figura 3.6: Exemplo da saída do processamento de um RDF pelo PAGE

3.5 Sumário

Neste capítulo foi apresentada um descrição pormenorizada da arquitectura do sistema HENDRIX, e dos módulos que o constituem: Minorthird, responsável pela extracção das entidades geográficas de textos; PAREDES, um módulo de software desenvolvido para associar as entidades geográficas extraídas com conceitos nas ontologias geográficas, desambiguação dos significados geográficos e geração de resumos geográficos; PAGE, um módulo para fazer a extracção de entidades mencionadas em grande escala com o HENDRIX, sobre um *cluster* suportado pelo HADOOP.

No próximo capítulo é apresentada a avaliação deste *software* no GikiCLEF 2009 e a sua aplicação na extracção de resumos geográficas da WPT05. As heurísticas de desambiguação são avaliadas com base em artigos da Wikipedia portuguesa.

Capítulo 4

Resultados

Neste capítulo apresentam-se resultados de várias avaliações do HENDRIX, nomeadamente do treino do modelo de reconhecimento de entidades geográficas (EG) com as Colecções Douradas do HAREM (Mota and Santos, 2008a), da sua utilização no Gikiclef edição de 2009, um evento de avaliação de sistemas de respostas a tópicos com um âmbito geográfico. Apresentam-se também os resultados da aplicação do HENDRIX à geração de resumos geográficos das páginas web da WPT05, uma recolha da web portuguesa e uma avaliação das heurísticas usadas para avaliação dos resumos gerados, com base em artigos da Wikipedia. Neste processo foi realizada a identificação automática da língua presente nos documentos que constituem a colecção, sendo o método utilizado também descrito em detalhe.

4.1 Treino do modelo de Reconhecimento de Entidades Geográficas

O modelo matemático de *Conditional Random Fields* (CRF) usado para extraír as entidades geográficas de textos foi gerado com base nas Colecções Douradas (CD) dos eventos HAREM, ambos descritos no Capítulo 2. Foi realizada uma análise estatística inicial das entidades marcadas nas CD.

Os gráficos das frequências acumuladas para o número de ocorrências de cada entidade, mostram que cerca de dez entidades únicas são responsáveis por quase 25% de todas as ocorrências de entidades geográficas nas CD. As figuras 4.1 4.2 e 4.3 apresentam os gráficos das curvas numa escala logarítmica, para as CD do Primeiro HAREM, Mini-HAREM e Segundo HAREM, respectivamente. No eixo das abcissas, as entidades estão ordenadas pelo logaritmo na base 10 da sua posição, numa tabela de frequências, o eixo das ordenadas representa o logaritmo na base 10 do número de ocorrências para cada entidade. Os gráficos apenas contêm as entidades pertencente à categoria ”LOCAL”, não incluindo as do tipo ”VIRTUAL”. As tabelas 4.1 4.2 e 4.3 listam as dez entidades mais frequentes.

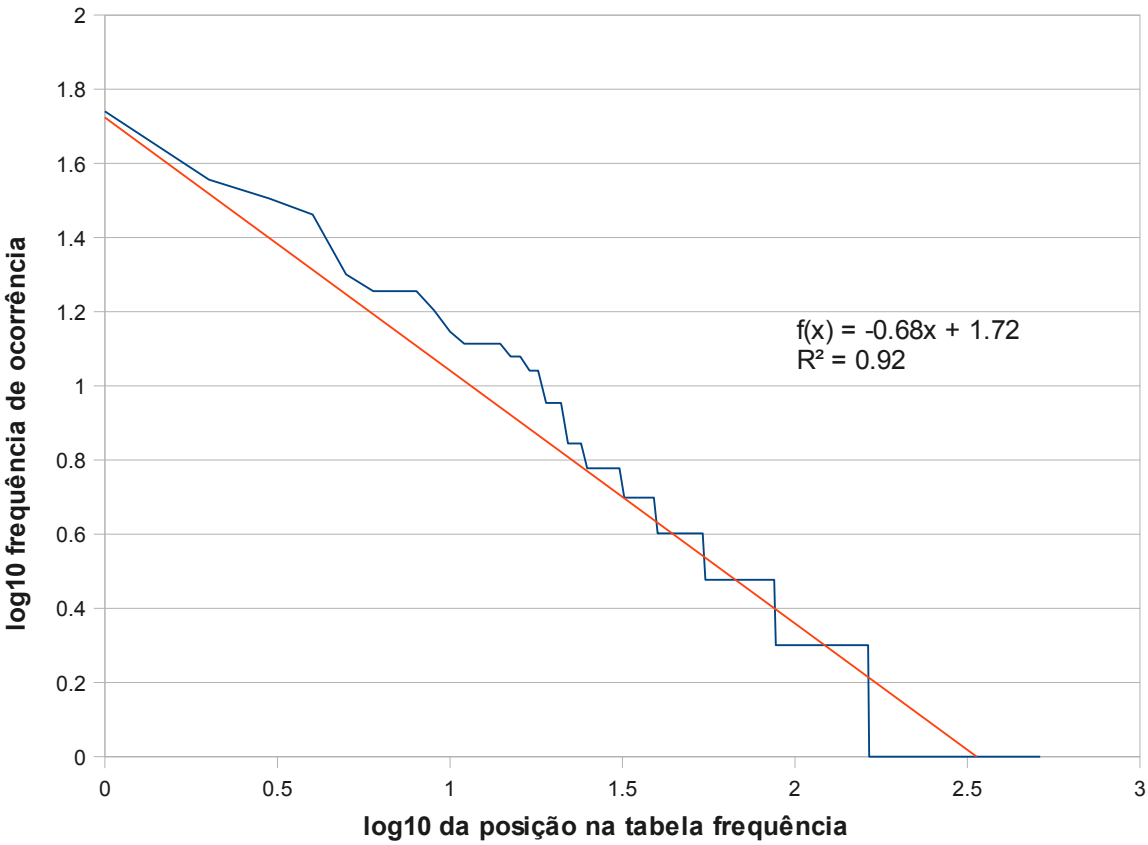

Figura 4.1: Ocorrências de EM geográficas na CD do HAREM I

Os gráficos mostram que as entidades da categoria ”LOCAL” seguem uma distribuição de acordo com a Lei de Zipf (Zipf, 1949), com um coeficiente de variação R^2 de cerca de 90%. Em cada colecção há um pequeno número de entidades responsáveis por uma grande parte de todas as ocorrências de entidades da categoria ”LOCAL” . Existe um número pequeno de entidades frequentes, e uma longa lista de entidades pouco frequentes.

As CD de cada uma das avaliações HAREM são disjuntas entre si, isto é, foram geradas a partir de documentos diferentes, mas têm no entanto entidades mencionadas (EM) comuns. Isto permite ao sistema aprender a extrair as mesmas entidades em contextos diferentes, podendo enriquecer as funções de característica geradas na fase de aprendizagem.

De maneira a ser possível fazer uma comparação do modelo de reconhecimento de entidades geográficas com outros sistemas desenvolvidos, utilizaram-se as CD do HAREM I e do Mini-HAREM para treinar o sistema, e as colecções do HAREM II para testar o modelo gerado. Assim foi possível comparar as métricas de Precisão, Abrangência e Medida-F com outros sistemas que tiveram uma participação na avaliação selectiva apenas para a categoria ”LOCAL” no HAREM II. A Tabela 4.4 apresenta uma comparação do

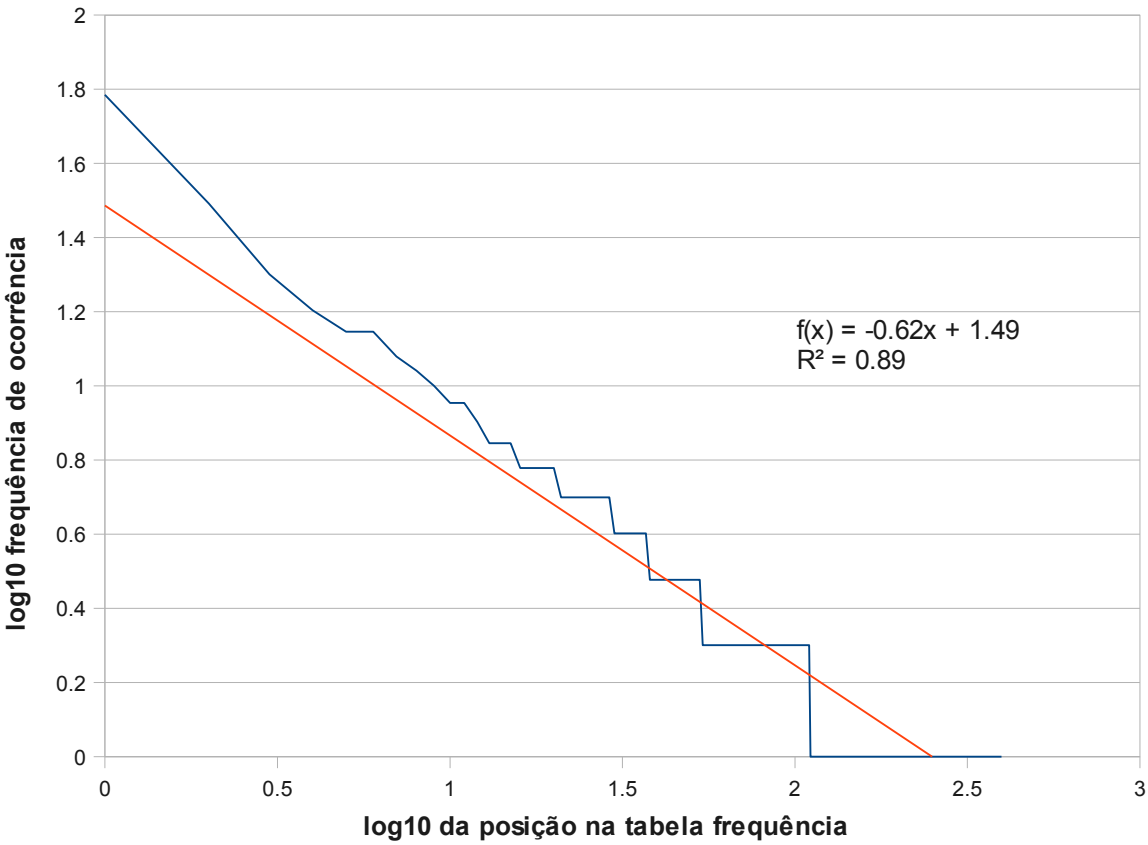

Figura 4.2: Ocorrências de EM geográficas na CD do Mini-HAREM

desempenho na identificação de EM do tipo LOCAL entre o Minorthird configurado para o uso de CRF (M3rd CRF), o componente de extracção de EM usado pelo HENDRIX, e outros sistemas.

Os outros sistemas apresentados utilizam para fazer reconhecimento de EM métodos

Entidade	Ocorrências	Freq. Acumulada
Brasil	55	4.80%
São Paulo	36	7.94%
Portugal	32	10.73%
Braga	29	13.26%
Lisboa	20	15.01%
Europa	18	16.58%
Porto	18	18.15%
Espanha	18	19.72%
Guimarães	16	21.12%
Marília	14	22.34%

Tabela 4.1: Entidades geográficas mais frequentes para a CD do HAREM I

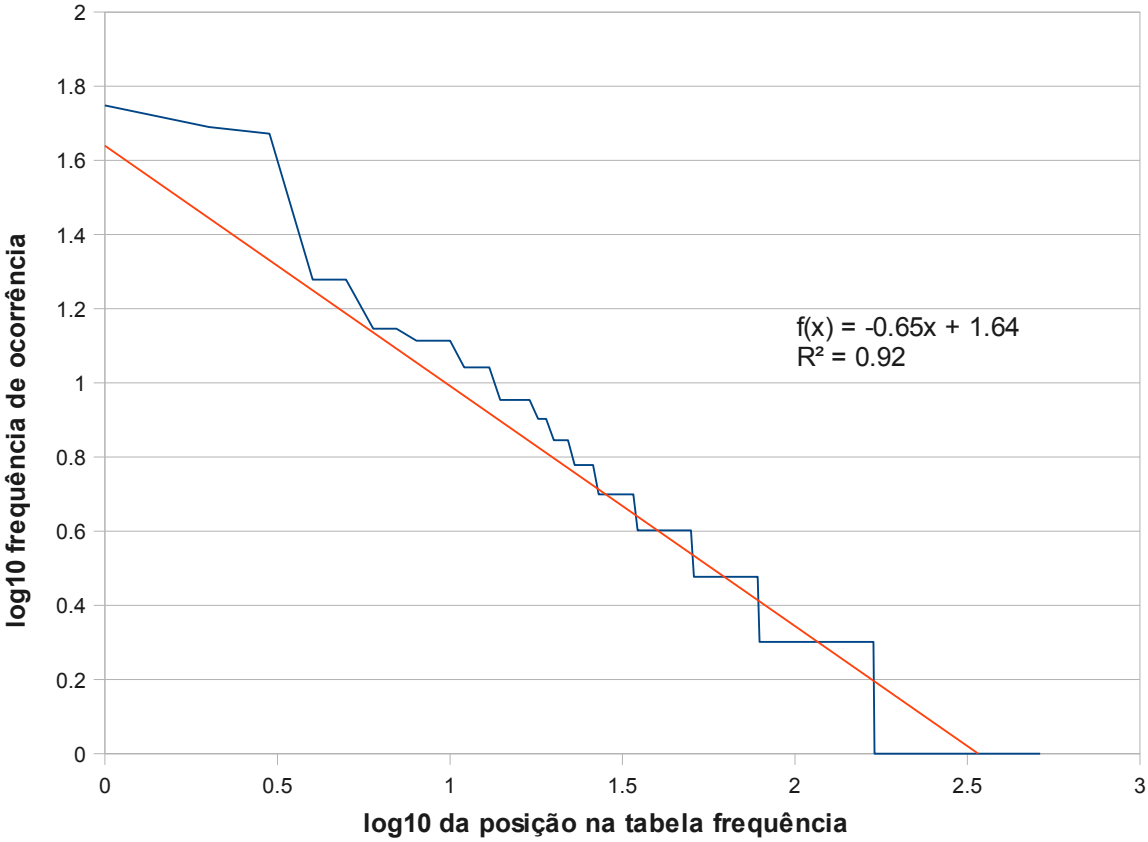

Figura 4.3: Ocorrências de EM geográficas na CD do HAREM II

linguísticos com regras definidas manualmente e em alguns casos a consulta de bases de conhecimento externas. O Minorthird por outro lado gerou as regras de detecção de forma automática, tendo por base textos em português anotados. Segue-se um descrição curta dos sistemas comparados:

Entidade	Ocorrências	Freq. Acumulada
Brasil	61	7.70%
São Paulo	31	11.62%
Itália	20	14.14%
Angola	16	16.16%
Braga	14	17.93%
Egito	14	19.70%
Portugal	12	21.21%
Santos	11	22.60%
São Vicente	10	23.86%
Europa	9	25.00%

Tabela 4.2: Entidades Geográficas mais frequentes para a CD do Mini-HAREM

O REMBRANDT (Cardoso, 2008) é um sistema de identificação e classificação de entidades mencionadas, tem ainda a possibilidade de fazer detecção de relações entre entidades mencionadas. Usa um conjunto de regras gramaticais para identificar a presença de uma EM e a Wikipedia como base de conhecimento externa para efectuar a classificação da EM identificada.

O SEIGeo (Chaves, 2008) foi desenvolvido especificamente para reconhecimento de entidades geográficas, aplica expressões que possam detectar a presença de uma entidade geográfica aos textos de modo a identificar possíveis entidades geográficas. Usa duas ontologias geográficas, a WGO e a Geo-Net-PT como dicionário de nomes para identificar entidades geográficas a partir das expressões e palavras extraídas.

O SeRELep (Bruckschen et al., 2008) é um sistema construído com o intuito de reconhecer relações entre as EM. As tarefas de identificação e classificação são realizadas por um analisador sintáctico, o PALAVRAS (Bick, 2000). No entanto a base para a tarefa de identificação é feita usando um método linguístico sem recorrer a bases de conhecimento externas, o que pode justificar, a alta abrangência e baixa precisão.

Comparando os resultados obtidos com os outros sistemas verifica-se que o desempenho do modelo de CRF é inferior a dois sistemas com que foi comparado. No entanto este ainda está longe de ser optimizado. Os resultados sugerem que através da geração de melhores funções de característica na fase de aprendizagem, quer através de uma CD melhorada ou através da sua codificação manual, os resultados de precisão e abrangência poderam aumentar.

O modelo gerado falha em muitos casos em reconhecer o significado de uma entidade consoante o seu contexto. Por exemplo, o termo "Portugal" é anotado como sendo da categoria LOCAL, quando se refere geograficamente ao país, em outros exemplos aparece anotado como ORGANIZACAO ou PESSOA, referindo-se ao governo de Portugal ou a uma entidade não geográfica. A CD usada para teste do modelo, foi também transformada

Entidade	Ocorrências	Freq. Acumulada
Lisboa	56	4.67%
Portugal	49	8.75%
Brasil	47	12.67%
Coimbra	19	14.25%
EUA	19	15.83%
Europa	14	17.00%
Porto	14	18.17%
França	13	19.25%
Detroit	13	20.33%
São Paulo	13	21.42%

Tabela 4.3: Entidades geográficas mais frequentes para a CD do HAREM II

	Precisão	Abrangência	Medida-F
REMBRANDT_3_corr	0,56	0,73	0,63
SEIGeo_2	0,71	0,51	0,59
HENDRIX (M3rd CRF)	0,64	0,45	0.53
SeRELeP_1	0,22	0,79	0,34

Tabela 4.4: Identificação de EM da categoria LOCAL no HAREM II

por forma a manter apenas as anotações para entidades geográficas, deixando o termo ”Portugal” não anotado quando empregue noutros contextos que não o geográfico. No entanto, os testes mostraram que o modelo extrai sempre ”Portugal”, isto é, identifica-o sempre como entidade geográfica, elevando assim o número de falsos positivos.

Com base nas CD do HAREM I e do Mini-HAREM foram geradas as funções de característica, estando cada uma associada às várias etiquetas de classificação. A Tabela 4.5 mostra o número de funções de característica geradas para cada uma das possíveis etiquetas de classificação. As funções de característica associadas a cada etiqueta são analisadas de seguida, sendo apresentadas para cada etiqueta as 10 com maior peso.

Para a etiqueta BEGIN, as 10 *features* com o maior peso são apresentadas na Tabela 4.6. As *features* testam, por exemplo, se o termo actual é igual a ”av” ou ”rua”, ”serra” ou ”vila”, se o termo tem a primeira letra em maiúscula, ou se o termo anterior foi classificado como NEG ou se é igual a ”em”.

As 10 funções de característica com maior peso associadas à etiqueta CONTINUE, que marca uma entidade a ser extraída constituída por mais do que um termo, são apresentadas na Tabela 4.7. As com maior peso testam se o termo anterior foi classificado com CONTINUE OU BEGIN. Outros exemplos de funções geradas são a de o termo ser um conector de substantivos como ”de” ou ”do”, a presença de termos à esquerda classificados com o valor de ”av” ou ”rua”.

A etiqueta END marca o final de uma entidade a ser extraída constituida por mais do que um termo, na tabela 4.8 estão as 10 *features* com maior peso associadas a esta etiqueta. Estas testam a etiqueta do termo à esquerda para o valor de BEGIN ou CONTINUE. Os termos à esquerda estarem etiquetadas como NEG e começarem com caracteres

Etiqueta	Nº *features*
BEGIN	1 862
CONTINUE	1 693
END	1 845
NEG	1 322 121
Total	1 327 521

Tabela 4.5: Distribuição das funções de característica pelas etiquetas de classificação

Funções geradas	Peso
tokens.eq.lc.av	5.05
previousLabel.1.NEG	4.65
right.token_0.eq.charTypePattern.X+x+	3.89
tokens.eq.lc.s	3.75
tokens.eq.lc.rua	3.75
left.tokenNeg_1.eq.lc.em	3.68
tokens.eq.lc.vila	3.44
tokens.eq.lc.belo	3.3
tokens.eq.lc.st	3.03
tokens.eq.lc.serra	2.8

Tabela 4.6: Funções de característica de maior peso associadas à etiqueta BEGIN

maiúsculos e terminarem com minúsculos ou poderem ser iguais a ou ”serra”, ”vila”.

UNIQUE é a etiqueta que marca entidades a extrair constituídas por apenas um único termo. A Tabela 4.9 mostra as 10 funções de maior peso geradas para esta etiqueta. A maior parte faz uma comparação com o valor do termo actual, sendo que alguns dos termos contidos coincidem com algumas das entidades mais frequentes nas CD como ”Itália”, ”Guimarães” ou ”Marília”, que fazem parte das entidades mais repetidas das CD usadas na fase de aprendizagem.

A etiqueta NEG, que marca termos que não são constituintes de entidades a reconhecer foi a etiqueta com mais funções associadas, com o maior peso estão as que analisam a etiqueta do termo anterior, como por exemplo ser também uma etiqueta NEG, UNIQUE ou END, ou ser *null*, o que indica o início de frase. A Tabela 4.10 mostra a 10 com o maior peso.

Funções geradas	Peso
previousLabel.1.localContinue	17.48
previousLabel.1.localBegin	16.9
left.tokenNeg_1.eq.lc.av	4.98
tokens.eq.lc.de	3.48
left.tokenNeg_1.eq.lc.s	3.36
tokens.eq.charTypePattern.x+	3.24
left.tokenNeg_2.eq.lc.estado	3.16
tokens.eq.lc.do	2.93
left.tokenNeg_2.eq.lc.av	2.6
left.tokenNeg_2.eq.lc.rua	2.54

Tabela 4.7: Funções de característica de maior peso associadas à etiqueta CONTINUE

Funçõe geradas	Peso
previousLabel.1.localBegin	20.36
previousLabel.1.localContinue	18.83
left.tokenNeg_2.eq.lc.s	3.78
left.tokenNeg_2.eq.charTypePattern.X+x+	3.44
left.tokenNeg_1.eq.lc.belo	3.35
left.tokenNeg_1.eq.lc.presidente	2.91
tokens.eq.lc.março	2.71
tokens.eq.lc.unidos	2.69
left.tokenNeg_1.eq.lc.serra	2.68
left.tokenNeg_1.eq.lc.st	2.67

Tabela 4.8: Funções de maior peso associadas à etiqueta END

Funções geradas	Peso
tokens.eq.lc.tortosendo	6.19
tokens.eq.lc.itália	6.13
tokens.eq.charTypePattern.X+x+	5.91
tokens.eq.lc.covilhã	5.88
previousLabel.1.NEG	5.88
tokens.eq.lc.guimarães	5.4
tokens.eq.lc.marília	5.32
tokens.eq.lc.andradas	5.32
tokens.eq.lc.araraquara	5.2
tokens.eq.lc.pisões	5.04

Tabela 4.9: Funções de característica de maior peso associadas à etiqueta UNIQUE

Funções geradas	Peso
previousLabel.1.NEG	19.38
previousLabel.1.localUnique	12.69
previousLabel.1.localEnd	9.3
previousLabel.1.null	8.56
tokens.eq.charTypePattern.x+çãx+	8.01
tokens.eq.charTypePattern.x+ãx+	7.8
tokens.eq.charTypePattern.x+íx+	6.9
tokens.eq.charTypePattern.x+	6.79
tokens.eq.lc.filosofia	6.38
tokens.eq.charTypePattern.x+çx+	6.04

Tabela 4.10: Funções de característica de maior peso associadas à etiqueta NEG

Entidade	Precisão	Abrangência	Medida-F
PESSOA	0.5915	0.4095	0.4840
LOCAL	0.4590	0.5006	0.4789
EVENTO	0.3281	0.2515	0.2847
ORGANIZAÇÃO	0.4464	0.4783	0.4618

Tabela 4.11: Resultados da avaliação do modelo para o GikiCLEF 2009

4.2 GikiCLEF

O GikiCLEF (Santos and Cabral, 2009) é um evento de avaliação de sistemas de respostas a perguntas. O seu objectivo é avaliar sistemas que encontram documentos ou artigos na Wikipedia contendo a resposta a uma determinada pergunta ou uma informação necessária. O processo de procura da resposta envolve raciocínio geográfico por parte dos sistemas. Os sistemas participantes têm que responder a um conjunto de tópicos, usando a Wikipedia como base de conhecimento, devolvem o título de um ou mais artigos onde está a resposta.

Cada sistema recebe um conjunto de tópicos em várias línguas, representando uma necessidade de obter determinada informação. Cada tópico representa uma pergunta válida e realista por parte de um utilizador. O sistema tem que depois produzir uma lista de respostas, em todas as línguas nas quais consegue encontrar repostas.

O GikiCLEF em 2009 abrangeu as seguintes línguas: Búlgaro, Neerlandês, Inglês, Alemão, Italiano, Norueguês nas vertentes Dano-Norueguês *(Bokmål)* e Novo-Norueguês *(Nynorsk)*, Português, Romeno e Espanhol. Os tópicos estavam assim traduzidos em 10 línguas diferentes.

O HENDRIX fez parte do sistema desenvolvido pelo XLDB para a participação na edição de 2009 do GikiCLEF. Foi criado para esta participação um modelo de CRF de forma a reconhecer não apenas lugares mas também organizações, eventos e pessoas. A fase de treino fez uso das CD do HAREM descritas na secção anterior. As CD do HAREM I e Mini-HAREM serviram de treino e as do HAREM II foram utilizadas para testar o modelo.

Foi usado apenas um modelo de CRF para extrair os quatro diferentes tipos de entidades descritos acima. Observou-se que o desempenho para a categoria LOCAL baixou ligeiramente, e muitas das entidades foram identificadas correctamente, mas classificadas com a categoria errada. Isto leva a crer que se deveria ter treinado um modelo em separado para cada uma das categorias, opção que poderia ter levado a melhores resultados. A Tabela 4.11 mostra os resultados da avaliação. Em (Cardoso et al., 2009) é feita uma descrição detalhada do processo de constriução do sistema e de melhoramentos a fazer no futuro.

O modelo treinado foi aplicado a um *dump* da Wikipedia portuguesa de 20 de Janeiro de 2009, de cada artigo foram extraídas as entidades mencionadas. Para os ar-

tigos com mais do que uma entidade LUGAR reconhecida foram extraídas as relações semânticas com base na ontologia Wiki WGO 2009. Desta forma foi criado um sumário para cada artigo, com as entidades detectadas e as relações entre os lugares, a Figura 4.4 mostra um exemplo para o artigo da Wikipedia sobre a cidade do Porto. O evento teve uma participação de 8 sistemas, ficando o sistema desenvolvido pelo XLDB na segunda posição.

```
Braga  part-of-> Portugal
Aveiro  part-of-> Portugal
Aveiro  part-of-> Aveiro
Porto  part-of-> Porto
Porto  part-of-> Portugal
Lisboa  part-of-> Portugal
Braga  part-of-> Braga
Forte de Sao Francisco Xavier do Queijo  part-of-> Porto
Vila Nova de Gaia  part-of-> Porto
Lisboa  part-of-> Lisboa
Guimaraes  part-of-> Braga

Antonio Cupertino (PESSOA)
Rui Veloso  (PESSOA)
Palacio da Bolsa  (PESSOA)
Patrimonio Mundial  (ORGANIZACAO)
Museu de Arte Contemporanea (ORGANIZACAO)
Ciencia Mundial (ORGANIZACAO)
D. Henrique (PESSOA)
Museu de Arte Sacra (ORGANIZACAO)
FC Porto  (ORGANIZACAO)
Dom Luis  (PESSOA)
Associacao Comercial do Porto (ORGANIZACAO)
D. Joao (PESSOA)
Ponte Dom Luis  (ORGANIZACAO)
D. Maria  (PESSOA)
Feira da Queima (ACONTECIMENTO)
```

Figura 4.4: Exemplo do sumário gerado pelo HENDRIX para o GikiCLEF 2009

4.3 Anotação da WPT05

A WPT-05 é uma recolha da web portuguesa, feita em 2005. Contem mais de 10 milhões de documentos da web portuguesa recolhida pelos batedores (*crawlers*) do motor de pesquisa Tumba! e produzida pelo Pólo XLDB da Linguateca (XLDB and Linguateca, 2006). Engloba conteúdos recolhidos de acordo com os seguintes critérios:

- alojados sob um domínio .pt
- escritos em português e alojados sob um domínio .com, .org, .net ou .tv, desde que tenham sido referenciados por um *links* de, pelo menos, uma página alojada sob um

domínio .pt.

É disponibilizada em duas versões:

- RDF/XML, que inclui os meta-dados, e o texto extraído dos conteúdos recolhidos.
- ARC (do Internet Archive), com os conteúdos armazenados tal como foram recolhidos

A versão RDF/XML da WPT05 tira partido da tecnologia RDF e da especificação OAI-ORE (http://www.openarchives.org/ore/) para a representação de duplicados e hierarquias entre páginas, apresentando os meta-dados de recolha e o texto extraído de cada URL. As suas características são:

- Sem textos duplicados. Os textos dos documentos marcados como duplicados não são incluídos, indicando-se apenas referência para o URL recolhido com esse texto.
- Preservação dos domínios. A relação de associação das páginas ao domínio de onde foram recolhidas é indicada nos meta-dados.
- Documentos ricos em texto. Os documentos incluídos são apenas os que têm um dos seguintes tipos MIME: application/pdf, application/postscript, application/vnd.ms-office, text/html, text/plain, text/rtf.
- Todos os ficheiros da colecção estão codificados em UTF-8.
- RDF/XML. Cada ficheiro da colecção é um ficheiro XML válido, possibilitando o seu manuseamento pelas ferramentas de software de tratamento de documentos em RDF e XML.

4.3.1 Identificação Linguística

A recolha feita incidiu em conteúdos publicados na web portuguesa, cujos documentos se encontram escritos nas mais variadas línguas. Como apenas se pode fazer reconhecimento de entidades mencionadas em português, em virtude de o modelo do HENDRIX só ter sido treinado com documentos em português, foi necessário identificar a língua de cada documento de forma a seleccionar os que seriam processados pelo HENDRIX.

Para aplicar a geração de resumos apenas aos documentos em português, aplicou-se uma técnica baseada em n-gramas de forma a classificar automaticamente a língua em que cada documento se encontra escrito (Cavnar et al., 1994). Esta técnica tinha sido já anteriormente aplicada a classificar uma recolha da web portuguesa (Martins and Silva, 2005).

A ideia base de categorizar um texto usando n-gramas, é a de calcular um perfil pertencente a uma categoria desconhecida e compará-la com uma lista de perfis de documentos

cuja categoria é conhecida. Os perfis são constituidos por uma lista das n-gramas mais frequentes num dado documento, ordenadas pela sua frequência As categorias que mais se aproximam da categoria desconhecida são dadas como o resultado da classificação. A Figura 4.5 mostra o processo de geração de perfis de categorização e de classificação linguística para um um dado texto.

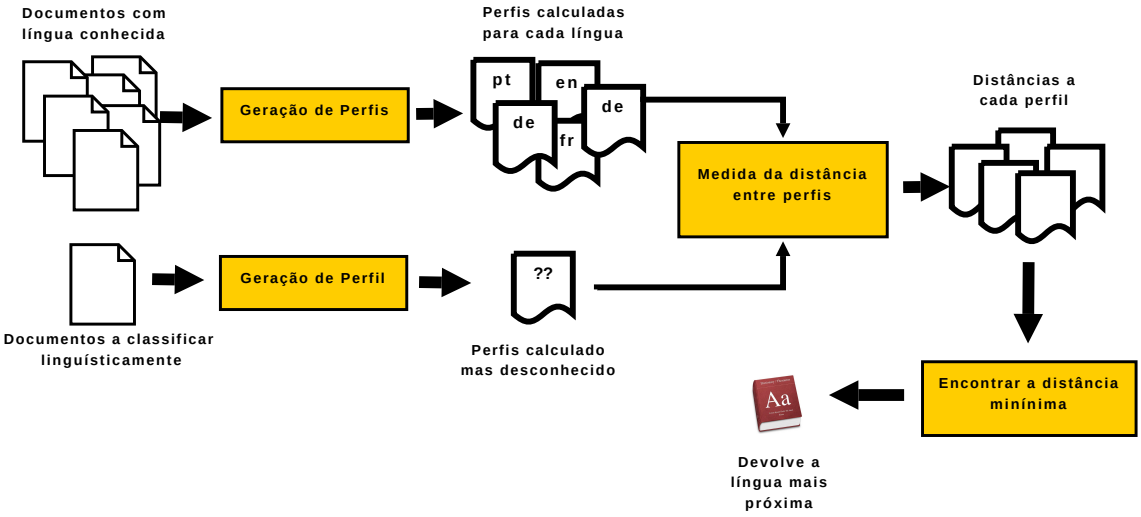

Figura 4.5: Classificação Linguística com base em n-gramas

Uma n-grama é um divisão em n-caracteres de uma dada cadeia de caracteres, onde o seu tamanho é maior que n. Por exemplo, a palavra ”HENDRIX” é composta pelas seguintes n-gramas :

Unigrams: _, h, e, n, d, r, i, x,_

Bigrams: _h, he, en, nd, dr, ri, ix, x_

Trigrams: _he, end, dri, ix_, x__

A Figura 4.6 mostra como são computadas as distâncias entre perfis de n-gramas. Calcula-se a soma das distâncias entre as posições de cada n-grama no perfil da categoria e no documento.

O processamento do texto utilizado n-gramas tem algumas vantagens. Não é necessário atomizar do texto em palavras, tem-se n-gramas de caracteres em vez de palavras como unidade de informação, eliminando-se assim a tarefa de reconhecer palavras. Ao fazer a atomização do texto usando n-grams de caracteres, cada cadeia de caracteres é decomposta em pequenas partes. Assim erros de ortografia tendem a afectar apenas um número reduzido dessas pequenas partes. Isto é importante, dado que os textos a classificar são recolhidos da web, onde a qualidade da escrita é muito variável.

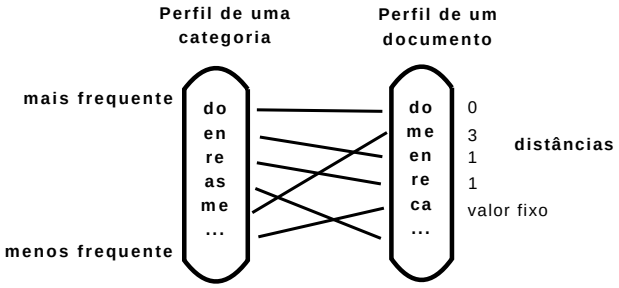

Figura 4.6: Distâncias entre dois perfis de n-gramas

NGramJ

Foi utilizado o *software* NGramJ (disponível em http://ngramj.sourceforge.net/) para identificar a língua em que cada documento se encontra escrito. O NGramJ implementa o método de n-gramas atrás descrito para classificar linguisticamente um documento. Contém perfis calculados para cerca 70 línguas, usando os n-gramas de tamanho até 4.

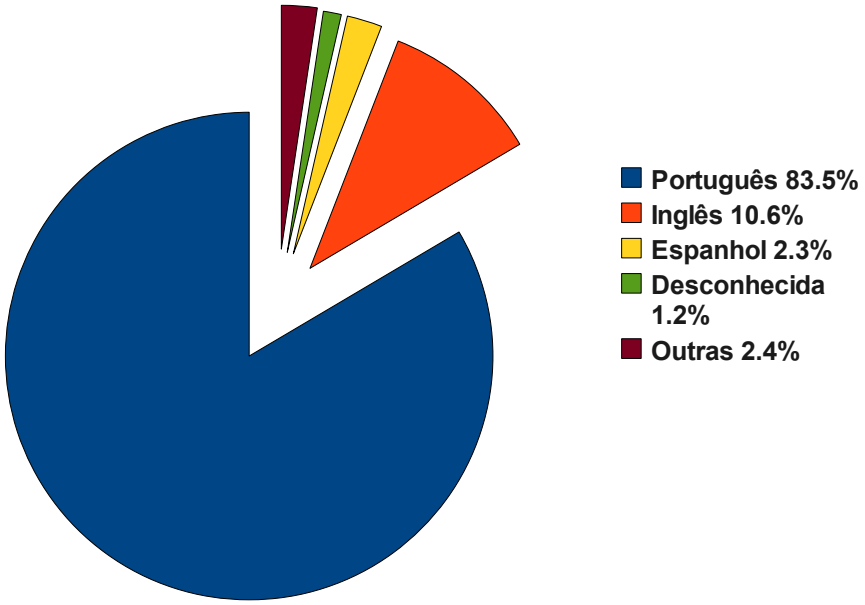

Figura 4.7: Línguas mais frequentes na WPT-05

A Figura 4.7 mostra o gráfico da distribuição das classificações linguísticas por documento, utilizando o software NGramJ e considerando apenas os documentos da WPT05 com mais de 200 bytes. Os documentos com classificação ”desconhecida” correspondem a recolhas de páginas onde grande parte do texto é constituindo por URL ou endereços de email, listagens de pastas num *webserver*, conteúdos para as quais não é possível identificar a língua em que foram escritos. Nessas, o perfil calculado fica muito distante

de qualquer um dos perfis categorizados. A distribuição é semelhante a uma anterior classificação de uma recolha da web portuguesa de 2003 (Martins and Silva, 2005). A Tabela 4.12 mostra a prevalências das 10 línguas mais frequentes. O Português, Inglês e Espanhol são as línguas mais encontradas na recolha. As variantes de português não são no entanto distinguidas, havendo muitos documentos escritos em português do Brasil.

Língua	Nº Documentos	Tamanho (Bytes)
Português	7.412.778 (59.19%)	25.906.873.629
Inglês	941.711 (7.52%)	3.589.560.517
Espanhol	206.732 (1.65%)	839.337.036
Desconhecida	106.195 (0.85%)	323.368.116
Outras	91.968(%0.73)	399.064.860
Alemão	63.073 (0.5%)	154.063.641
Francês	54.973 (0.44%)	202.188.956
Documentos $<$ 200 bytes	606.059 (4.84%)	-
sem texto	3.039.621 (24.247%)	-
Total	12.523.110 (100%)	-

Tabela 4.12: Classificação Linguística da WPT-05

4.3.2 Marcação de Entidades Geográficas Mencionadas

Os documentos que constituem a WPT05 foram submetidos ao módulo de *software* PAGE de modo a ser efectuada a extracção de entidades geográficas. Foi usado o modelo de CRF descrito na secção 4.1 treinado com as CD do HAREM I e MiniHarem e as ontologias geográficas: Geo-Net-PT, World Ontology Geography, Wiki Geo 2009, descritas no capítulo 2. O *cluster* de unidades de processamento foi constituído por 2 servidores, usando 10 unidades de processamento (*cores*):

- 4 x Intel(R) Xeon(R) CPU @ 2.50GHz
- 6 x Quad-Core AMD Opteron(tm) Processor 2350 @ 1GHz

A extracção de entidades geográficas sob esta configuração durou aproximadamente 16 dias. Dos cerca de 7,5 milhões documentos em português foram extraídas no total 78 326 entidades únicas. Depois de processadas pelo PAREDES foram encontradas correspondências nas ontologias para 18 586 (23.73%) das entidades encontradas.

A Tabela 4.13 mostra qual o número de entidades para as quais foram encontradas pelo menos um conceito numa ontologia. (Nota: A mesma entidade poderá estar em mais do que uma ontologia e mais do que um domínio dentro da mesma ontologia.)

A Geo-Net-PT foi a ontologia na qual mais correspondências com as entidades extraídas foram encontradas, seguida da Wiki WGO 2009, o que significa que grande parte

Ontologia	Nº Entidades	Percentagem
Geo-Net-PT 2.0	13 097	70.47%
Administrativo Sem *feature type* associado	8 175	-
Administrativo Com *feature type* associado	4 889	-
Físico Sem *feature type* associado	2 033	-
Físico Com *feature type* associado	189	-
World Geographic Ontology	2 191	11.79%
Domínio Administrativo	2 094	-
Domínio Físico	146	-
Wiki WGO 2009	8 742	47.04%

Tabela 4.13: Entidades extraídas com correspondências nas ontologias

das entidades extraídas dos textos têm âmbito no território português. No entanto algumas das entidades extraídas entidades têm um âmbito geográfico fora do território português.

Os textos que fazem parte dos documentos recolhidos na WPT05 têm origens diferentes, desde *blogs* pessoais, jornais, *sites* institucionais, de comércio on-line, entre outros, havendo diversos tipos de textos. Os nomes de entidades geográficas presentes nos textos são muitas vezes referidos por outros nomes que não os nomes oficiais, presentes no domínio administrativo da Geo-Net-PT.

Noutros casos, muitos dos nomes extraídos contêm erros ortográficos ou abreviaturas, o que sugere que seja usado um outro método para procurar representações na Geo-Net-PT. A Tabela 4.14 apresenta alguns exemplos.

Foram feitas apenas comparações de cadeias de caracteres, de mode que, para que seja retornado um conceito geográfico da Geo-Net-PT, a entidade extraída tem que conter os mesmo caracteres que o nome da sua representação. Um outro método, baseado em distância de edição (Levenshtein, 1966) é necessário para conseguir extrair o maior número de referências possíveis.

Outro problema encontrado, que reforça a ideia de usar esta técnica para comparação de caracteres, está relacionado com o uso de artigos definidos compostos nos nomes das entidades geográficas. Com bastante frequência encontram-se casos em que nos textos são extraídos nomes sem os artigos definidos mas que se encontram na Geo-Net-PT com os artigos (a Tabela 4.15 mostra alguns exemplos).

Algumas das entidades extraídas, são de facto geograficamente relevantes mas a sua geo-codificação não é possível na Geo-Net-PT. Existem entidades extraídas que designam locais ou zonas por outros nomes de que não os nomes administrativos oficiais, ou representam mais do que um local através de outros nomes, por exemplo:

"Área Metropolitana do Porto" referindo-se a uma área que agrupa 16 concelhos

"Castelo Lisboa" referência ao Castelo de São Jorge em Lisboa

Entidade extraída	Nome na Geo-Net-PT
Alavalde	Alvalade
Amadoa	Amadora
Av. da sidade de Aveiro	Avenida da Cidade de Aveiro
Avenida do Brazil	Avenida do Brasil
Avenida Dão Nuno Álvares Pereira	Avenida Dom Nuno Álvares Pereira
Caldas da Rinha	Caldas da Rainha
Caldas da Raínha	Caldas da Rainha
Castanheira do Ribetejo	Castanheira do Ribatejo
Cova da Mora	Cova da Moura
Figueró da Granja	Figueiró da Granja
Herdade do Zmbujal	Herdade do Zambujal
Jardim Botãnico	Jardim Botânico
Vila Franca do Canpo	Vila Franca do Campo
Vila Nova da Barquiinha	Vila Nova da Barquinha
Vila Nova da Gaia	Vila Nova de Gaia
Vila Nova de Famalcão	Vila Nova de Famalicão
Campo de Sant'Ana	Campo de Santana
Vila Nova de Mil Fontes	Vila Nova de Milfontes

Tabela 4.14: Erros ortográficos em entidades extraídas

"Baixa Portuense" ou "Baixa da Invicta" designação dada à zona central da cidade do Porto

"Baixa de Coimbra" designação dada à zona central da cidade de Coimbra

O modelo treinado conseguiu também extrair algumas moradas completas, como mostra a Tabela 4.16. Este tipo de expressões pode ser muito útil no processo de desambiguação já que além do tipo de entidade geográfica, indicam também a cidade a que esta pertençe.

Há conceitos geográficos que têm como nome datas, normalmente associadas a eventos de importância histórica, os textos das CD que fizeram parte do conjunto de aprendizagem contêm entidades geográficas com estas propriedades, o que leva o modelo CRF a extrair datas como locais. Foram extraídas 227 datas únicas sem nenhuma descrição

Entidade extraída	Nome na Geo-Net-PT
Av. 25 Abril	Avenida 25 de Abril
Av. Fontes Pereira Melo	Avenida Fontes Pereira de Melo
Av. Antonio Augusto Aguiar	Avenida António Augusto de Aguiar
Av. Fernão Magalhães	Avenida Fernão de Magalhães
Vale Milhaços	Vale de Milhaços

Tabela 4.15: Exemplos de falta de artigos definidos em EG extraídas da WPT05

Bonnertalweg 53129 Bona - Alemanha
Bouço 4820 Arões S. Romão
Rua General Bruce, 230 São Cristóvão
Rua Governador Mata nº 36
Rua General Humberto Delgado 4760 Vila Nova de Famalicão
Rua General Humberto Delgado 7160 Bencatel
Rua Humberto Delgado 2985-213 Pegões Velhos
Rua Oliveira Júnior, Nº 25 3700 São João da Madeira
Rua Padre Castilho 5150 Vila Nova de Foz Côa
Av. Francisco Pinto Pacheco 2670 Santo António Cavaleiros
Av. de Bordeaux 33850 Léognan - França

Tabela 4.16: Exemplos de moradas extraídas da WPT-05

```
'^[0-9][0-9]?\sde\s(\<Janeiro\>|\<Fevereiro\>|\<Março\>|
\<Abril\>|\<Maio\>|\<Junho\>|\<Julho\>|\<Agosto\>|
\<Setembro\>|\<Outubro\>|\<Novembro\>|
\<Dezembro\>).*'
```

Figura 4.8: Expressão regular utilizada para detectar datas

de tipo de conceito geográfico associado, como ”Avenida” ou ”Rua” o que sugere que o modelo treinado extrai datas ao ter aprendido a associar entidades geográficas a expressões de datas. A expressão regular na Figura 4.8 foi usada para identificar datas entre as entidades extraídas.

4.4 Avaliação e Âmbitos Geográficos

As heurísticas propostas para gerar os resumos geográficos, descritos na secção 3.3.2 foram avaliadas com base em artigos da Wikipedia portuguesa, referentes aos distritos portugueses. Foram seleccionados artigos da Wikipedia referentes a cada capital de distrito portuguesa, a Tabela 4.17 mostra para cada artigo, o tamanho do texto, o número de entidades extraídas pelo modelo CRF apresentado anteriormente e o número de entidades extraídas emparelhadas na Geo-Net-PT.

A Tabela 4.18 mostra os resultados para a primeira heurística usada para calcular o âmbito geográfico dos artigos da Wikipedia sobre capital de distrito. Esta heurística tenta calcular o âmbito geográfico do documento apenas através das relações entre as referências encontradas na Geo-Net-PT.

São extraídas as relações entre todas as referências de forma a construir um grafo. O âmbito geográfico é depois dado pela referência que mais arcos tem com outras referências, ou seja mais relações tens com outras referências encontradas. Esta heurística tem a vantagem de ser bastante rápida.

Capital de Distrito	Tamanho Textos (Kbytes)	Entidades Extraídas	Entidades com correspondentes Geo-Net-PT
Aveiro	6,1	33	24
Beja	12	9	6
Braga	72	125	71
Bragança	1,9	5	3
Castelo Branco	5,4	11	6
Coimbra	24	35	29
Évora	2,8	7	6
Faro	9,8	22	18
Guarda	11	24	17
Leiria	17	38	33
Lisboa	33	62	42
Portalegre	27	48	35
Porto	26	56	31
Santarém	4,7	15	11
Setúbal	12	27	17
Viana do Castelo	2,4	16	11
Vila Real	20	57	35
Viseu	31	71	40

Tabela 4.17: Entidades extraídas para os artigos da Wikipedia

Capital de Distrito	Tempo	Âmbito Geográfico
Aveiro	0m28.104s	Aveiro (Distrito)
Beja	0m9.765s	Beja (Distrito)
Bragança	0m1.722s	Norte (NT2)
Braga	2m39.845s	Norte (NT2)
Castelo Branco	0m5.157s	Beira Baixa (Província)
Coimbra	0m30.201s	Porto (Distrito)
Évora	0m3.970s	Alentejo Central (NT3)
Faro	0m16.924s	Algarve (NT2)
Guarda	0m21.143s	Guarda (Distrito)
Leiria	1m3.229s	Leiria (Distrito)
Lisboa	0m31.673s	Lisboa (NT2)
Portalegre	0m34.058s	Norte (NT2)
Porto	0m42.013s	Braga (Distrito)
Santarém	0m16.450s	Santarém (Distrito)
Setúbal	0m16.981s	Lisboa (NT2)
Viana do Castelo	0m18.375s	Viana do Castelo (Distrito)
Vila Real	1m6.765s	Minho (Província)
Viseu	1m19.529s	Norte (NT2)

Tabela 4.18: Avaliação da Heurística 1

Capital de Distrito	Tempo	Âmbito Geográfico (Pai Comum)
Aveiro	39m56.402s	Portugal (PAI)
Beja	2m51.158s	Continente (NT1)
Braga	54m24.442s	Continente (NT1)
Bragança	0m20.873s	Norte (NT2)
Castelo Branco	0m56.673s	Continente (NT1)
Coimbra	14m0.788s	Continente (NT1)
Évora	3m0.648s	Continente (NT1)
Faro	7m56.371s	Continente (NT1)
Guarda	14m26.773s	Continente (NT1)
Leiria	14m41.674s	Continente (NT1)
Lisboa	18m30.573s	Continente (NT1)
Portalegre	21m22.517s	Continente (NT1)
Porto	24m15.857s	Continente (NT1)
Santarém	2m45.902s	Continente (NT1)
Setúbal	4m44.003s	Continente (NT1)
Viana do Castelo	11m45.932s	Norte (NT2)
Vila Real	29m32.073s	Continente (NT1)
Viseu	39m28.529s	Continente (NT1)

Tabela 4.19: Avaliação da Heurística 2

Dos 18 artigos sobre as capitais de distrito de Portugal, para 6 o âmbito geográfico coincidiu com o distrito da qual a cidade é capital.

Para 10 dos artigos o âmbito geográfico calculado abrange uma área maior do que o distrito. Os artigos sobre Bragança, Braga, Portalegre, Viseu foram todos classificados como pertencente ao Norte usando a Nomenclatura Comum das Unidades Territoriais Estatísticas (NUT) de nível II. O mesmo aconteceu com os artigos de sobre Lisboa e Setúbal, ficando com o âmbito de Lisboa NUT de nível II. Também com os artigos de Évora e Faro, sendo-lhes atribuido o âmbito de Alentejo Central NUT de nível III e Algarve NUT de nível II, respectivamente.

Os artigos sobre Castelo Branco e Vila Real, ficaram também classificados com um âmbito geográfico superior ao distrito mas mais específico, neste caso províncias, Beira Baixa e Minho respectivamente. Dois artigos foram classificados com um âmbito geográfico mais específico, mas errado, os artigos de Coimbra e Porto.

A Tabela 4.19 mostra os resultados da avaliação da segunda heurística baseada em medidas de semelhança e descrita no secção 3.3.2

Alguns problemas ocorreram com na geração do âmbito geográfico, que levaram a ter um âmbito muito mais geral do que realmente o documento tem. Muitos dos artigos sobre as capitais de distrito contêm entidades que referenciam outros distritos ou concelhos.

Capital de Distrito	Tempo	Âmbito Geográfico (Relações)
Aveiro	42m32.536s	Aveiro (Distrito)
Beja	2m51.594s	Beja (Distrito)
Braga	49m24.954s	Norte (NT2)
Bragança	0m20.455s	Norte (NT2)
Castelo Branco	0m49.370s	Beira Baixa (Província)
Coimbra	15m30.650s	Coimbra (Distrito)
Évora	3m8.401s	Alentejo Central (NT3)
Faro	7m37.448s	Algarve (Província)
Guarda	13m45.904s	Guarda (Distrito)
Leiria	14m7.433s	Beira Litoral (Província)
Lisboa	16m37.365s	Grande Lisboa (NT3)
Portalegre	20m16.159s	Norte (NT2)
Porto	23m58.968s	Porto (Distrito)
Santarém	2m34.583s	Alentejo (NT2)
Setúbal	4m23.933s	Lisboa (NT2)
Viana do Castelo	11m16.082s	Viana do castelo (DST)
Vila Real	29m26.061s	Vila Real (Distrito)
Viseu	39m20.529s	Viseu (Distrito)

Tabela 4.20: Avaliação da Heurística 3

Identificador	Nome de Entidade	Tipo de Conceito
3945	Beja	DST
100	Cuba	CON
191	Mértola	CON
174908	Campismo	PAR
107896	Santo amaro	LOC

Tabela 4.21: Referências extraídas para o artigo sobre Beja

Por exemplo ao serem extraídos os nomes dos distritos ou províncias vizinhas, e as suas respectivas referências correspondentes, fica-se com referências muito dispersas no grafo que corresponde às relações na Geo-Net-PT. Ao tentar calcular o âmbito usando o antecessor comum entre as referências encontradas, chega-se na maior parte das vezes Portugal Continental NUT 1, a raiz do grafo em forma de árvore invertida, que corresponde à Geo-Net-PT.

Outro problema, prende-se com falsos positivos, isto é, por exemplo, nomes de pessoas ou monumentos e que são erradamente emparelhadas a conceitos geográficos na ontologia. Em muitos casos as referências estão muito distantes no grafo das outras referências encontradas, levando também à observação do problema anterior, ao calcular o âmbito, o único pai em comum é Portugal Continental NUT 1. Por exemplo, no artigo sobre a capital de distrito Beja, ao serem emparelhadas as entidades extraídas e depois de aplicar as heurísticas de redução e as medidas de semelhança entre as referências das entidades extraídas, ficam apenas as referências apresentadas na Tabela 4.21

Neste caso, ”Campismo” e ”Santo Amaro” são dois falsos positivos, fora do Distrito de Beja. A referência a ”Santo Amaro” corresponde uma localidade pertencente ao Distrito de Coimbra e ”Campismo” é referência que corresponde a um parque na Nazaré, no Distrito de Leiria. Ao calcular o pai comum mais próximo entre estas referências chega-se a Portugal Continental.

No entanto, aplicando a terceira heurística, definida na Secção 3.3.2, que em vez do antecessor comum, extrai de todas as relações a entidade que mais relações agrega, os resultados são diferentes, como mostra a Tabela 4.20.

Desta maneira, no exemplo fica-se com, Beja (Distrito), e com com Cuba (Concelho) e Mértola (Concelho) com a relação $filho - de$ com Beja (Distrito), eliminado os falsos positivos. Esta heurística falhou para apenas um dos artigos calculando o âmbito geográfico de Santarém como Alentejo (NT2).

4.5 Conclusão

As anotações para entidades geográficas nas Colecções Douradas do HAREM seguem, como foi mostrado, a Lei de Zipf. Um pequeno número de entidades únicas são res-

ponsáveis por quase 25% de todas as ocorrências de entidades geográficas nas CD. Para gerar um modelo de CRF com melhores funções característica é necessário a sua codificação manual ou dados de treino anotados tendo em conta a tarefa específica de extracção de informação geográfica e não a tarefa geral de Reconhecimento de Entidades Mencionadas (REM).

A terceira heurística foi a que melhor resultados teve na inferência do âmbito geográfico. No entanto o uso de medidas de semelhança na fase de desambiguação é um processo bastante demorado, devido ao número elevado de consultas que são feitas à base dados, é necessário carregar os dados usados nestes cálculos para memória de forma a acelerar o processo.

A geração de resumos para a WPT05 está em curso, devido a ser um processo de computação longa, não foi possível inclui-os neste capítulo.

Capítulo 5

Conclusão e Trabalho Futuro

O trabalho apresentado nesta dissertação teve como objectivo o desenvolvimento de um sistema a executar num *cluster* de computadores para extracção de entidades geográficas de documentos utilizando técnicas de aprendizagem automática, nomeadamente os *Conditional Random Fields* (CRF), e de seguida, a associação das entidades extraídas a conceitos numa ontologia geográfica e posterior geração de resumos geográficos que caracterizam o âmbito geográfico dos documentos.

O sistema HENDRIX foi desenvolvido para este fim, é constituído por diferentes componentes: o modelo de CRF do Minorthird, treinado com as Colecções Douradas do HAREM, responsável pela extracção de entidades geográficas dos textos; o PAREDES é o módulo de *software* responsável pela associação de entidades a conceitos geográficos; o PAGE é módulo de software para extracção de entidades usando um *cluster* baseado no Hadoop. O sistema consulta ontologias geográficas para validar e associar as entidades geográficas extraídas.

O modelo de CRF gerado foi treinado com as Colecções Douradas do evento HAREM I e MiniHAREM e depois testado com a Colecção Dourada do HAREM II, tendo 64% de Precisão, 45% de Abrangência. Este foi o modelo usado para extrair entidades geográficas da WPT05.

Os documentos da WPT05, uma recolha da Web portuguesa, foram identificados linguisticamente, dos identificados como escritos em português, cerca de 7 500 000 num total de 26 Gbytes de texto, foram processados pelo sistema HENDRIX.

Foram extraídas 78 326 entidades únicas, das quais 18 586 (23,73%) correspondem a conceitos geográficos. Das entidades que representam pelo menos um conceito geográfico, 13 097 (70,47%) estão na Geo-Net-PT, uma ontologia com âmbito no território português.

O processo de geração dos resumos está em curso, não havendo ainda estatísticas finais.

5.1 Experiências com o modelo CRF

O modelo de CRF gerado precisa de ser melhorado, os valores das medidas de precisão e abrangência, calculadas com base nas Colecções Douradas do HAREM II dão um desempenho com uma classificação inferior a outros sistemas baseados apenas em regras codificadas manualmente. No entanto esses sistemas foram desenvolvidos tendo em consideração o evento HAREM. O modelo aqui treinado apenas gerou as regras com base em textos anotados.

O evento HAREM foi criado para avaliar a tarefa geral de Reconhecimento de Entidades Mencionadas, as Colecções Douradas foram anotados com esse propósito e não a extracção de informação geográfica, tarefa à qual se pretende chegar com o treino do CRF. As anotações das CD falham muitas vezes em captar o contexto geográfico de uma entidade. Existem entidades anotadas da categoria local que não incluem o seu tipo de entidade geográfica, por exemplo:

```
..das lojas Modelo de <LOCAL>Eiras</LOCAL>, no distrito de <LOCAL>
    Coimbra</LOCAL> e de <LOCAL>Lagoa</LOCAL>, no concelho de <LOCAL>
    Portimao</LOCAL>
```

O tamanho das CD e os diferentes exemplos de anotações existentes não são suficientes para gerar funções de característica capazes de captar toda a informação geográfica num documento. Os testes feitos com artigos das capitais de distrito da Wikipedia mostram que o modelo falha em extrair entidades geográficas de expressões como:

"O município é limitado a norte pelos municípios de Cuba e Vidigueira, a leste por Serpa, a sul por Mértola e Castro Verde e a oeste por Aljustrel e Ferreira do Alentejo."

As funções de característica geradas não são suficientes para captar as expressões de localização deste exemplo, usando pontos cardeais antes das referências aos concelhos, sendo que a única entidade geográfica a ser extraída é "Cuba".

Delboni (2005) faz um estudo sobre expressões de localização para o português como fonte de conceitos geográficos em páginas Web. Reune um conjunto de expressões que indicam a presença de uma entidade geográfica. A codificação dessas expressões em funções de característica pode levar a um aumento do desempenho o modelo na extracção de informação geográfica.

5.2 Inferência de Âmbitos Geográficos

Das três heurísticas desenvolvidas para inferir o âmbito geográfico de um documento, duas mostraram bons resultados, a primeira e terceira.

A primeira que não aplica medidas de semelhança semântica. O grafo gerado pela primeira heurística tem um processo de desambiguação menos rigoroso em relação às outras duas, apenas extrair as relações de todas as referências encontradas para as entidades extraídas. Ao extrair as relações algumas referências ficam fora do grafo, por não terem relação com nenhuma das outras referências. Por outro lado, poderá haver casos onde para a mesma entidade extraída, existem no grafo mais do que uma referência. No entanto, e tendo em consideração que o resultado final é chegar a um âmbito geográfico, esta técnica apresentou bons resultados. O âmbito geográfico é dado pela referência que mais relações tem com as outras.

A segunda heurística aplica medidas de semelhança entre os pares de entidades, pela ordem que ocorrem no texto. De seguida calcula o antecessor comum mais próximo, usando-o para definir o âmbito geográfico do documento. Esta técnica de gerar o âmbito geográfico do documento mostrou-se pouco eficaz, estes ficam com uma área geográfica muito superior à esperada.

A terceira heurística aplicada usa medidas de semelhança semântica que permitem chegar a um grafo muito mais pequeno e específico, contendo apenas uma referência para cada entidade desambiguada. As medidas de semelhança são aplicadas a cada par de entidades extraídas, pela sua posição no texto. O processo é mais demoroso, mas ao final fica-se apenas com uma referência única para cada entidade extraída. Desse conjunto são extraídas todas as relações possíveis, sendo o âmbito geográfico dado pela referência que mais relações com as outras.

A primeira e terceira heurística, mesmo com um modelo de abrangência baixa, tendo em conta a quantidade de informação existente nos artigos geraram âmbitos geográficos muito próximos dos esperados.

5.3 Conclusões

A falta de dados de treino específicos para a tarefa de extracção de informação geográfica e a consequente geração de funções de característica baseadas apenas nos dados de treino existentes, as Colecções Douradas do HAREM, deram origem a um modelo pouco abrangente. Um modelo gerado com melhores dados de treino e possivelmente também com algumas funções de característica codificados à mão poderia levar a uma extracção mais rigorosa de entidades geográficas, com o tipo de entidade geográfica associado para reduzir o número de referências devolvidas nas consultas às ontologias.

Um problema verificado é a ocorrência de falsos positivos no processo de identificação, o modelo extrai nomes de entidades, que referenciam no texto nomes de pessoas ou outras categorias que não local. Estes falsos positivos podem propagar-se para o processo de classificação, na geo-codificação, muitas vezes estas entidades erradamente extraídas

como locais são emparelhadas em referências geográficas. Isto leva a que entrem referências erradas para o processo de desambiguação e consequentemente levem à inferência de um âmbito geográfico errado.

É necessário um melhor conjunto de dados de treino, contendo exemplos positivos e negativos, de forma a forçar a aprendizagem de quando um nome está ou não num contexto geográfico. Um modelo assim treinado poderia baixar o número de entidades extraídas que podem corresponder tanto a conceitos geográficos como a outras classes de entidades. Aprendendo a contextualizar através de exemplos, o modelo poderia distinguir entre contextos, quando a entidade representa o nome de uma pessoa e quando representa uma localidade.

Das entidades extraídas da WPT05 a partir deste modelo, 23,73% correspondem a conceitos geográficos nas ontologias, havendo no conjunto de entidades não emparelhadas uma parte que corresponde a informação geográfica. No entanto o processamento desta informação de forma a emparelhar com conceitos nas ontologias não é trivial, porque as entidades contêm erros ortográficos ou porque se referem a conceitos presentes na ontologia, mas com um nome diferente ou porque seguem padrões para os quais o *software* por mim desenvolvido não captura, como moradas por exemplo. Há também informação geográfica extraída mas que não se encontra nas ontologias usadas. No entanto, a sua quantidade é difícil de contabilizar, seria necessária uma verificação por humanos, ou como alternativa recorrer a outras ontologias mais ricas.

A ontologia com âmbito no território português, a Geo-Net-PT, foi onde a maior parte das entidades foram emparelhadas, mostrando que das entidades geográficas extraídas dos documentos escritos em português 70,47% têm âmbito no território Português.

Os sistema HENDRIX, é flexível no sentido em que facilmente se integra um novo modelo de CRF, bastando recorrer ao Minorthird para treinar um novo modelo, e usando esse modelo no PAGE para extrair entidades. Poderá eventualmente vir a ser usado para fazer extracção de outro tipo de entidades para colecções de documentos.

5.4 Trabalho futuro

Um elemento essencial para melhorar o desempenho do sistema HENDRIX é um corpus de treino anotado com o objectivo de extracção de informação geográfica. É necessário reanotar as CD tendo em conta a tarefa de informação geográfica e a geração de funções de característica do modelo CRF. Vimos que os artigos da Wikipedia referentes às capitais de distrito contêm muitas referências geográficas e descrições com recurso a expressões de localização. Estes artigos podem também ser anotados e usados na fase de treino do modelo.

Obtendo funções de característica mais específicas, por exemplo, que testem a entidade a extrair para nomes associadas a tipos de conceitos geográficos, contidos na ontolo-

gia, por exemplo: ”distrito”, ”concelho”, ”freguesia”. No fundo juntando a aprendizagem por regras à que foi feita com base exclusivamente em exemplos.

Isto poderia levar o modelo a extrair mais entidades geográficas, incluindo também os seus tipos de conceitos associados. Uma codificação destas expressões em funções de característica poderia aumentar o desempenho do sistema. Novas funções poderam são geradas através da reanotação dos textos, ou através de codificação manual. Delboni (2005) faz um estudo de expressões de posicionamento para o português do Brasil a partir de textos recolhidos na web, a codificação destas expressões em funções de característica poderá trazer melhores resultados.

Existe outro tipo de entidades nos textos que poderam dar evidências de uma geografia associada ao documento, tais como: organizações - instituições, universidades, empresas ou associações - através da localização das suas sedes ou filiais; eventos, através do sítio onde ocorreram.

Treinando modelos diferentes, usando as CD transformadas, cada um deles específico para uma classe de entidades, poderíamos depois procurar extrair entidades dos textos que denotam eventos e organizações. Utilizando um modelo CRF para cada tipo de entidade a extrair, poderá levar a melhores resultados. As entidades extraídas poderam depois ser usadas em consultas a bases de conhecimento externas, como a DBpedia (Bizer et al., 2009) para fazer o emparelhamento e extracção de relações.

O módulo de *software* PAREDES não faz tratamento de moradas, no entanto Borges et al. (2007) fazem um estudo de endereços e moradas presentes em páginas web, no futuro os padrões descritos nesse trabalho deverão ser incorporados no PAREDES de forma a conseguir captar estas expressões.

As técnicas de desambiguação usando as medidas de semelhança poderam ser exploradas de outra foram. Rauch et al. (2003) mostram que há uma correlação geo-espacial alta entre entidades geográficas que estão próximas num texto. Em alternativa a calcular a medida de semelhança entre pares de entidades geográficas, pela ordem que estas ocorrem no texto, pode-se tentar fazer o cálculo entre pares mais próximo no texto.

Outras formas de gerar um âmbito geográfico passam pela utilização de coordenadas geográficas. A Geo-Net-PT inclui coordenadas geográficas para Freguesias, Concelhos e Distritos. Usando as coordenadas geográficas que representam centróides, caixas delimitadoras ou polígonos, pode-se tentar encontrar a região mais pequena que é capaz de englobar o maior número de referências. Leidner et al. (2003) usam esta técnica para fazer desambiguação de topónimos, uma adaptação poderia ser usada para seleccionar qual o âmbito de um documento.

Embora tivesse sido feito o emparelhamento com conceitos do domínio físico da Geo-Net-PT, as suas referências, e as relações inter-domínio, entre os domínios físicos e administrativos não foram exploradas Estas relações além de enriquecer o resumo geográfico, podem ajudar no processo de desambiguação e inferência do âmbito geográfico.

Bibliografia

Lada A. Adamic. Zipf, power-law, pareto - a ranking tutorial. Technical report, Information Dynamics Lab, HP Labs, HP Labs, Palo Alto, CA 94304, October 2000. URL `http://www.hpl.hp.com/research/idl/papers/ranking/`.

B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining, 2002. URL `citeseer.ist.psu.edu/berendt02towards.html`.

Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. May 2001.

Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University, Denmark, November 2000.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, July 2009. ISSN 15708268. doi: 10.1016/j.websem.2009.07.002. URL `http://dx.doi.org/10.1016/j.websem.2009.07.002`.

Karla A. V. Borges, Alberto H. F. Laender, Claudia B. Medeiros, and Clodoveu A. Davis, Jr. Discovering geographic locations in web pages using urban addresses. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 31–36, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-828-2.

Mírian Bruckschen, José Guilherme Camargo de Souza, and Renata Vieira e Sandro Rigo. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. 2008. URL `http://www.linguateca.pt/HAREM/actas/Capitulo_14-MotaSantos2008.pdf`.

Nuno Cardoso. Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In *Encontro do Segundo HAREM, PROPOR 2008*, Aveiro, Portugal, 7 de Setembro 2008.

Nuno Cardoso. Avaliação de sistemas de reconhecimento de entidades mencionadas. Master's thesis, Faculdade de Engenharia da Universidade do Porto, December 2006.

Nuno Cardoso, Mário J. Silva, and Diana Santos. Handling Implicit Geographic Evidence for Geographic IR. In *Proceedings of the 17th Conference on Information and Knowledge Management, CIKM'2008*, Napa Valley, CA, EUA, 27–29 de Outubro 2008. ACM. accepted for publication.

Nuno Cardoso, David Batista, Francisco J. Lopez-Pellicer, and Mário J. Silva. Where in the wikipedia is that answer? the xldb at the gikiclef 2009 task. In Carol Peters et al, editor, *Working Notes of CLEF 2009*, Corfu, Greece, October 2009.

William Cavnar, , William B. Cavnar, and John M. Trenkle. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

Marcirio Chaves and Diana Santos. What kinds of geographical information are there in the portuguese web? In *PROPOR 06 - 7th Workshop on Computational Processing of Written and Spoken Language*, number 3960 in LNCS, Itatiaia, Rio de Janeiro, Brasil, May 2006. Springer.

Marcirio Chaves, Mário J. Silva, and Bruno Martins. A geographic knowledge base for semantic web applications. In *20th Brazilian Symposium on Databases - SBBD*, pages 40–54, Uberlândia, Minas Gerais, Brazil, October 2005.

Marcirio Chaves, Catarina Rodrigues, and Mário J. Silva. Data model for geographic ontologies generation. In *XATA 2007 - XML: Aplicações e Tecnologias Associadas*, February 2007.

Marcirio Silveira Chaves. Geo-ontologias e padrões para reconhecimento de locais em textos: a participação do sei-geo no segundo harem. In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM.*, 2008.

Marcirio Silveira Chaves. *Uma Metodologia para Construção de Geo-Ontologias*. PhD thesis, Faculty of Sciences, University of Lisbon, September 2009. URL `http://www.linguateca.pt/documentos/TeseDoutMarcirioChaves2009.pdf`.

William W. Cohen. Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data, http:// minorthird.sourceforge.net. 2004. URL `http://minorthird.sourceforge.net`.

William W. Cohen. Open archives initiative object reuse and exchange (oai-ore). URL `http://www.openarchives.org/ore/`.

H. Cunningham. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics, 2nd Edition*, 2005.

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. pages 137–150, 2004. URL `http://www.usenix.org/events/osdi04/tech/dean.html`.

Tiago Marques Delboni. Expressões de posicionamento como fonte de contexto geográfico na web. Master's thesis, Universidade Federal de Minas Gerais de Belo Horizonte, Agosto 2005.

Ian Densham and James Reid. A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 79–80, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Ronen Feldman. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521836573.

William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 233–237, Morristown, NJ, USA, 1992. Association for Computational Linguistics. ISBN 1-55860-272-0.

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. *SIGOPS Oper. Syst. Rev.*, 37(5):29–43, December 2003. ISSN 0163-5980. doi: 10.1145/1165389.945450.

Felix Jungermann. Named entity recognition mit conditional random fields. Master's thesis, Universität Dortmund, 7 2006.

Roman Klinger and Katrin Tomanek. Classical Probabilistic Models and Conditional Random Fields. Technical Report TR07-2-013, Department of Computer Science, Dortmund University of Technology, December 2007.

John D. Lafferty, Andrew Mccallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 31–38, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, February 1966. URL http://adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=1966SPhD...10..707L.

Bruno Martins. *Geographically Aware Web Text Mining*. PhD thesis, Faculty of Sciences, University of Lisbon, August 2008.

Bruno Martins and Mário J. Silva. A statistical study of the tumba! corpus. In *Advances in Natural Language Processing, 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004, Proceedings*, pages 384–394, 2004. Also available as University of Lisbon, Faculty of Sciences, Technical Report DI/FCUL TR 4-4.

Bruno Martins and Mário J. Silva. Language identification in web pages. In *ACM-SAC-DE, 20th ACM Symposium on Applied Computing, Document Engeneering Track*, pages 764–768, April 2005.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

Cristina Mota and Diana Santos. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. 2008a. URL http://www.linguateca.pt/LivroSegundoHAREM/.

Cristina Mota and Diana Santos. *Apêndice A. Segundo HAREM: Directivas de anotação*. 2008b. URL http://www.linguateca.pt/HAREM/actas/Apendice_A-MotaSantos2008.pdf.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

Erik Rauch, Michael Bukatin, and Kenneth Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 50–54, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Mark Sanderson and Janet Kohler. Analyzing geographic queries, 2004.

Diana Santos. Caminhos percorridos no mapa da portuguesificação: A linguateca em perspectiva. 2009. URL http://www.linguateca.pt/Diana/download/Santos2009Linguamatica.pdf.

Diana Santos and Luís Miguel Cabral. Gikiclef: Crosscultural issues in an international setting: asking non-english-centered questions to wikipedia, to appear. In *Working Notes of CLEF 2009*, Corfu, Greece, 30 September-2 October 2009.

Diana Santos and Nuno Cardoso. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. FCCN, 2008. URL `http://www.linguateca.pt/aval_conjunta/LivroHAREM/`.

Mário J. Silva, Bruno Martins, Marcirio Chaves, Nuno Cardoso, and Ana Paula Afonso. Adding geographic scopes to web resources. *CEUS - Computers, Environment and Urban Systems*, 30(4):378–399, July 2006.

W3C. Rdf vocabulary description language 1.0: Rdf schema, 2004.

W3C. Sparql query language for rdf, 2008.

XLDB and Linguateca. A wpt 05 é um recurso criado pela equipa de investigação xldb do lasige (http://xldb.di.fc.ul.pt/) em conjunto com a linguateca (http://www.linguateca.pt/). 2006.

George Kingsley Zipf. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Reading, MA, 1949.