Gestão de Sistemas e Computação

Reconhecimento de Entidades Mencionadas Geográficas e Suas Relações em Textos em Português

Trabalho final para Laboratório de GSC Aplicado

Elaborado por António Elias Sílvio Monteiro

Estudante nº 20070966

Orientador: Prof. Doutor Marcirio Silveira Chaves

Barcarena, Julho de 2010

Universidade Atlântica

Gestão de Sistemas e Computação

Reconhecimento de Entidades Mencionadas Geográficas e Suas Relações em Textos em Português

Trabalho final para Laboratório de GSC Aplicado

Elaborado por António Elias Sílvio Monteiro

Estudante nº 20070966

Orientador: Prof. Doutor Marcirio Silveira Chaves

Barcarena, Julho de 2010

O autor é o único responsável pelas ideias expressas neste relatório.

Reconhecimento de Entidades Mencionadas Geográficas e Suas Relações em Textos em Português
Gestão de Sistemas e Computação

Agradecimentos

Muitos foram aqueles e aquelas que durante este meu percurso académico contribuíram para que pudesse chegar ao fim desta etapa. A todos e todas o meu muito obrigado por esse apoio, mas quero agradecer em especial a:

- À Paula, minha mulher, pela força emocional demonstrada, pelo apoio incondicional sobretudo nos piores momentos, pela paciência, pelo amor;
- Aos meus pais e sogros pela serenidade demonstrada, pelo apoio emocional e incondicional, por acreditarem em mim;
- À Universidade Atlântica, por me proporcionar um ensino de qualidade;
- À Câmara Municipal de Oeiras pelo financiamento da minha licenciatura;
- Ao Professor Marcirio Chaves por me ter desafiado a encontrar novos caminhos com este trabalho, por me ter orientado e me ter levado a bom porto;
- À turma de Gestão de Sistemas e Computação, pela camaradagem e amizade que criamos, entrámos no curso como perfeitos estranhos e saímos deixando saudades;
- Ao Ricardo Ramalho, meu colega de trabalho, pelo apoio dado principalmente em épocas de frequências e de trabalhos de grupo, "aguentando" o serviço sozinho;
- Ao Miguel Faria, meu colega de trabalho, pela ajuda e apoio sempre disponível;
- A todos os professores e professoras que me deram aulas, com todos aprendi algo que certamente irei aplicar na minha vida futura;
- Aos meus gatos, companheiros incondicionais nos muitos dias e noites de estudo;

Abstract

Natural language automatic recognition is a complex problem. Although it is well solved to English language, in Portuguese language it is on its early steps which make it important to develop algorithms to deal with Portuguese.

This work describes information extraction techniques using raw texts to recognize named geographic entities and their relationships. These techniques are implemented in a system, REMPT which takes a raw text as input and returns the same text tagged with recognized entities and their relations if they exist.

Key-Words: Geographic Entities, Internet, Information Extraction.

Resumo

O reconhecimento automático de linguagem natural é um problema complexo. Apesar de estar bastante desenvolvido na língua inglesa, na língua portuguesa ainda está numa fase inicial pelo que se torna importante o desenvolvimento de algoritmos capazes de lidar com o português.

Este trabalho apresenta técnicas de extracção de informação em textos não estruturados para o reconhecimento de entidades mencionadas geográficas e suas relações. Essas técnicas são implementadas num sistema, O REMPT, que recebe como entrada um texto não estruturado e retorna o mesmo texto anotado com as entidades reconhecidas e as relações existentes entre as mesmas, caso existam.

Palavras-Chave: Entidades Geográficas, Internet, Extracção de Informação

Índice

Agradecimentos I
Abstract II
Resumo III
Índice IV
Ìndice de Tabelas V
Índice de Figuras V
1. Introdução 1
2. Conceitos e Trabalhos Relacionados 3
2.1 Conceitos 3
2.1.1 Recuperação de Informação 3
2.1.2 Extracção de Informação 3
2.1.3 Ontologia 5
2.1.4 Web Semântica 5
2.1.5 A Língua Portuguesa 6
2.2 Trabalhos Relacionados 7
2.2.1 Avaliação de Sistemas de REM: Os eventos HAREM 7
2.2.2 O sistema REMBRANDT 9
2.2.3 O sistema CaGE 10
2.2.4 O sistema SEI-Geo 13
2.2.5 O sistema SeRELeP 15
3. O sistema REMPT 17
3.1 Requisitos e Arquitectura de Software 18
3.1.1 Requisitos 18
3.1.2 Arquitectura de software 20
3.2 Arquitectura do sistema REMPT 21
3.2.1 Pré-Processamento 23
3.2.2 Identificação e Classificação de Entidades Geográficas 28
3.2.3 Reconhecimento de Relações 33
3.2.4 Saída de Resultados 35
3.3 Experiências 37
3.3.1 Experiência 1 37
3.3.2 Experiência 2 38
3.4 REMPT e os Trabalhos Relacionados 39
4. Considerações finais 41
4.1 Conclusões 41
4.2 Limitações 42
4.3 Trabalhos Futuros 43
Bibliografia 44
Anexo A 47
Tecnologias Utilizadas 48
RDF 48
OWL 50
SPARQL 51
SemWeb.NET 52

Virtuoso 52

Índice de Figuras

Fig.1-Exemplo de Construção de um arbusto 14
Fig 2-Arquitectura de software do REMPT 20
Fig.3-Arquitectura Geral do sistema REMPT 22
Fig.4 – Algoritmo para encontrar entidades multi-palavra 25
Fig 5- Um exemplo de texto processado pelo REMPT 36
Fig.6 –Exemplo de um elemento RDF 49
Fig.7 – Exemplo de um elemento N3 49

Índice de Tabelas

Tabela 1 -Utilizadores mundiais da internet e estatística da população (Stats, 2010) 6
Tabela 2- Descrição quantitativa da Geo-Net-PT01 (Chaves, 2009). 24
Tabela 3 -Expressões de contexto asociadas a referênicas geográficas (Martins, 2007) 31
Tabela 4 - Expressões adicionadas 31
Tabela 5 – Resultado da pesquisa por “Lisboa” usando SPARQL 51

1. Introdução

Actualmente os meios de acesso à informação são bastante variados, desde os mais tradicionais, como os jornais, as revistas ou a televisão até ao mais moderno, a Internet.

A Internet, apesar de ter sido idealizada para um projecto militar permite também uma forma de comunicação global e rápida. O mesmo tipo de informação, que antes era veiculada lentamente, é agora transmitido de forma quase instantânea para todo o mundo em simultâneo. As agências de notícias, por exemplo, conseguem enviar imagens para a redacção de um jornal no momento em que essas imagens são captadas, usando simplesmente um dispositivo com acesso à Internet. Quando queremos investigar sobre qualquer assunto podemos recorrer à pesquisa na Internet, usando motores de busca ou outros meios que se encontrem disponíveis. É possível também divulgar a nossa vida e nossa forma de pensar, em redes sociais, abrindo, assim a porta a novos meios de socialização. Podemos, por isso, afirmar que a Internet é a maior fonte de informação disponível hoje em dia. Apesar dessa quantidade de informação, os sistemas possuem limitações, tais como a dificuldade em conseguir obter informação relevante e com qualidade para determinadas pesquisas, como por exemplo, as pesquisas efectuadas em linguagem natural.

Um exemplo destas limitações pode ser encontrado ao usarmos o motor de resposta a perguntas em inglês, Answers.com. Se fizermos a pergunta “Where is Lisbon?” a primeira resposta será a página, em inglês, da Wikipedia sobre Lisboa. No entanto, se fizermos uma pergunta mais específica como “Where is Rua Augusta in Lisbon?” este motor de busca retorna uma lista de resultados provenientes da pesquisa feita no Google, ou seja, uma lista de páginas que não responde à pergunta. Portanto, ainda falham.

Apesar da noção da dificuldade que existe na língua portuguesa em conseguir relacionar entidades mencionadas em textos com locais existentes geograficamente, é

possível identificar e classificar locais através do uso de técnicas de extracção de informação (EI) e do uso de heurísticas.

Este trabalho apresenta o tratamento de informação geográfica em textos em português de modo a suportar sistemas de recuperação de informação e resposta a perguntas.

Este trabalho tem como objectivos principais:

Descrever técnicas de extracção de informação usadas para reconhecer locais geográficos em textos não estruturados;

Desenvolver um software que permita a entrada de textos não estruturados e o reconhecimento de entidades geográficas e suas relações presentes nesses textos. Este software basear-se-á na análise do texto, reconhecendo expressões que possam identificar locais usando para isso um dicionário de termos pré-definidos e uma geo-ontologia desenvolvida no âmbito de um dos trabalhos relacionados, para identificar as relações existentes entre os locais reconhecidos.

Este trabalho está estruturado como segue:

- O capítulo 2 descreve os conceitos subjacentes a este trabalho como por exemplo os conceitos de extracção e de recuperação de informação assim como trabalhos relacionados com esta matéria.

- O capítulo 3 começa por descrever em detalhe a forma como foi funciona o sistema de reconhecimento de entidades mencionadas geográficas, REMPT, mostra a sua arquitectura de software e a arquitectura geral de funcionamento do sistema, descrevendo em detalhe cada fase do processamento de texto.

- Neste capítulo são também explicadas as experiências efectuadas, os problemas encontrados e as soluções implementadas.

- O capítulo 4 descreve as conclusões do trabalho e as limitações encontradas, terminando com o que está planeado para trabalhos futuros.

2. Conceitos e Trabalhos Relacionados

Neste capítulo explico conceitos subjacentes ao tema deste trabalho, bem como os principais Sistemas de Reconhecimento de Entidades Mencionadas Geográficas.

Antes de entrarmos nos conceitos que dizem respeito às entidades geográficas propriamente ditas, convém explicar o que se entende por entidades mencionadas no contexto deste trabalho.

O termo Entidade Mencionada (EM) é uma tradução livre do mesmo conceito usado em inglês, *named entiy* e que significa numa tradução literal, entidades com nome próprio. Este é o termo adaptado pela Linguateca (Linguateca, 2007) para reconhecer EM que podem ser de várias categorias, como Pessoa, Organização ou Evento.

2.1 Conceitos

2.1.1 Recuperação de Informação

Os sistemas que usam a técnica de recuperação de informação (RI) têm como objectivo principal a disponibilização de documentos de acordo com a "força" das palavras-chave usadas pelo utilizador. A Wikipedia define a RI como sendo uma ciência para a procura de documentos e de metadados sobre documentos, incluindo a procura em bases de dados e na Web (Wikipedia, 2010a). Os sistemas de RI mais comuns e que todos os utilizadores da Internet usam, são os motores de busca, como por exemplo o Google, o Yahoo ou mais recentemente, o Bing. Actualmente estes motores de busca deixaram de ser simplesmente um sítio onde procurar informação "arquivada" na Web, e passaram a misturar esse conceito de RI com o conceito de extracção de informação. Deste modo, hoje é possível fazermos pesquisas que nos oferecem como resposta informação com mais qualidade do que a que víamos no início destes motores de busca.

2.1.2 Extracção de Informação

Os sistemas de extracção de informação (EI) têm como objectivo principal, localizar informações específicas, dentro de um documento, pedidas em linguagem

natural. Para (Cowie & Wilks, 2000) a extracção de informação pode ser definida como o processo que selectivamente estrutura e combina dados encontrados explicitamente ou implicitamente em um ou mais textos.

Imaginemos o seguinte cenário:

Numa empresa multinacional, um colaborador X pretende reunir-se com os colaboradores Y e Z, que estão, geograficamente separados, em vários pontos do mundo. O X pede ao seu sistema para lhe dizer qual será o melhor dia, dentro da semana de trabalho, em que se possa reunir com Y e Z. O sistema irá procurar na agenda dos três colaboradores e verificar qual é o melhor dia, hora e local, devido à diferença dos fusos horários, em que se podem reunir e, irá marcar essa reunião na agenda de cada um, avisando-os em seguida. Caso algum dos colaboradores tenha dificuldade no dia e hora marcada, pode individualmente dizer ao sistema para recalcular para outro dia.

Este cenário já é real e faz parte de um projecto maior, desenvolvido pela CISCO, em Portugal no âmbito do projecto a que chamam de *Telepresença* (http://www.cisco.com/web/BR/solucoes/tele_index.html) e encaixa-se no conceito de Web Semântica (WS) segundo o descrito por (Berners-Lee et al., 2001), que afirma que para termos uma WS é necessário dar aos computadores acesso a colecções estruturadas de informação que lhes permita inferir acerca de qualquer assunto e assim tomar, automaticamente uma decisão ponderada, de acordo com regras estabelecidas. Esta é apenas uma das muitas aplicações práticas que podem ser feitas com as tecnologias que estão a ser usadas para sistemas de EI, nomeadamente, o uso de ontologias que são baseadas em XML e em RDF (Ver Secção 2.1.3). De acordo com (Cowie & Wilks, 2000), para que esta tecnologia seja útil há que ter a habilidade de produzir este tipo de sistemas sem termos de recorrer a todos os recursos existentes para o Processamento de Linguagem Natural (PLN).

(Desnsham & Reid, 2003) dividiram o processo de extracção de informação geográfica em duas etapas: *geo-parsing* e *geo-coding*. O *geo-parsing* diz respeito à

identificação dos locais enquanto o *geo-coding* diz respeito à classificação desses locais e à sua desambiguação.

Derivadas dessas técnicas, surgiram os sistemas de resposta a perguntas que se baseiam sobretudo na EI para poderem dar uma resposta mais precisa ao utilizador.

Em Portugal, o uso de técnicas de EI ainda está no seu inicio, não existindo, ainda, um sistema de resposta a perguntas como os que já existem para a língua inglesa. Embora o meu trabalho seja no âmbito do reconhecimento de entidades geográficas e as suas relações, as técnicas usadas poderão servir para, termos um motor de busca ou um sistema de resposta a perguntas que reconheça essas entidades, na língua portuguesa.

2.1.3 Ontologia

Existem várias definições para o termo ontologia, para este trabalho interessa a que diz respeito à Ciência da Computação e, nesse domínio, uma ontologia é definida como sendo um "modelo de dados que representa um conjunto de conceitos, dentro de um domínio e os relacionamentos entre eles" (Wikipedia, 2010e). Algumas das áreas onde são usadas ontologias são a Inteligência Artificial, a Engenharia de Software, a arquitectura de informação e, mais recentemente, na implementação de modelos para a Web Semântica.

2.1.4 Web Semântica

Em termos linguísticos, a semântica refere-se ao estudo do significado, em todos os sentidos do termo, no entanto, em termos computacionais, a chamada Web Semântica (WS) é definida como sendo uma extensão da Web actual a qual possibilita a interacção e cooperação entre computadores e humanos. A WS interliga significados de palavras e, neste âmbito, tem como finalidade conseguir atribuir um significado (sentido) aos conteúdos publicados na Internet de modo que seja perceptível tanto pelo humano como pelo computador. A WS tem como objectivo o desenvolvimento de tecnologias e linguagens que tornem a informação legível para as máquinas. O uso de linguagens e tecnologias como *eXtensible Markup Language* (XML), *Resource*

Description Framework (RDF) ou ontologias, entre outras, favorecerá o aparecimento de serviços Web que garantam interoperabilidade e cooperação.

2.1.5 A Língua Portuguesa

Sendo considerada a 7ª língua mais falada no mundo, por cerca de 150 milhões de pessoas (Wikipedia, 2009), é bastante importante que nos debrucemos sobre a forma de, através de meios automáticos, conseguirmos extrair informação relevante de textos, que respondam a perguntas concretas. O problema que pretendo abordar neste trabalho prende-se precisamente com a dificuldade que existe em interpretar a nossa língua, de forma automática, seja pela sintaxe vasta que é usada, pelas formas gramaticais ou pelas expressões que se usam. Para tratar o problema da interpretação automática de textos em português, resolvi focar-me especificamente nas entidades mencionadas geográficas e suas relações.

De acordo com o site http://www.Internetworldstats.com, o número de pessoas, em 2009, estimadas como sendo utilizadoras da Internet eram cerca de 6.767.805.208 distribuídas da forma como se pode ver na Tabela 1:

Regiões do mundo	População· (2009)	Utilizadores da Internet Dez. 31, 2000	Últimos dados de utilizadores de Internet	Penetração· (% População)	Crescimento 2000-2009(%)
África	991.002.342	4.514.400	**86.217.900**	8.7	1.809.8
Ásia	3.808.070.503	114.304.000	**764.435.900**	20.1	568.8
Europa	803.850,858	105.096,093	**425.773.571**	53.0	305.1
Médio Oriente	202.687,005	3.284.800	**58.309.546**	28.8	1.675.1
América do Norte	340.831.831	108.096.800	**259.561.000**	76.2	140.1
América Latina	586.662.468	18.068.919	**186.922.050**	31.9	934.5
Austrália	34.700.201	7.620.480	**21.110.490**	60.8	177.0
TOTAL MUNDIAL	6.767.805.208	360.985.492	**1.802.330.457**	**26.6**	399.3

Tabela 1 - Utilizadores mundiais da Internet e estatística da população (Stats, 2010).

Dado este número de pessoas, ninguém sabe ao certo qual a quantidade de páginas existentes na Internet, mas de acordo com uma notícia publicada pelo Blog oficial do Google (Google, 2008) eram cerca de 1 trilião (10^{12}).

Estes números servem para termos uma noção da quantidade de informação que temos à disposição de cada um de nós e, no entanto, não conseguimos na maioria das vezes fazer uma pergunta e obter uma resposta concreta. O problema dessa informação, bastante valiosa é encontrar-se na sua maioria numa forma não estruturada, existindo apenas em textos dentro de páginas Web.

Imaginemos o seguinte cenário que hoje em dia é real: as grandes empresas e organismos públicos contratam empresas para pesquisarem e recolherem notícias, a que se chama *clipping*, em que apareçam notícias que dizem respeito a essas empresas ou organismos. Actualmente, essa recolha é na sua maioria feita manualmente sendo que as empresas pesquisam nos motores de busca normais tal como o Google, o Bing ou o Yahoo por notícias relacionadas com cada empresa cliente. Algumas empresas já usam outros métodos de procura, tendo aplicações feitas à medida para permitir outro tipo de buscas. Se houvesse um sistema de EI suficientemente rápido, estas empresas encontrariam talvez menos informação mas com muito mais qualidade.

É para resolver este tipo de problemas, encontrar o que queremos, que em meados da década de 70 o conceito de EI, que no seu sentido mais lato consiste no processo de interpretação de informação não estruturada de forma a que, obedecendo ao assunto em causa, retorne a informação de que necessitamos.

2.2 Trabalhos Relacionados

Nesta secção apresento dos trabalhos relacionados sobre EI e extracção de relações (ER), que serviram de base para a construção do sistema REMPT, descrito no capítulo 3.

2.2.1 Avaliação de Sistemas de REM: Os eventos HAREM

Em 1987 com o surgimento das *Message Understanding Conference* (MUC, 1987), cujo objectivo era melhorar a forma como se realizava a EI num texto não estruturado, surgiu a preocupação de se especializar essa EI de modo a canalizar os esforços para determinadas áreas de actuação. Essas áreas incluem a EI em textos relacionados com manobras militares de navios, passando pela identificação de

actividades terroristas na América Latina ou pela análise de acidentes de aviação. No entanto, essas conferências serviram o seu propósito apenas para a língua Inglesa. Em Portugal, em 2006 surgiu, através da Linguateca (http://linguateca.pt), o HAREM.

Este trabalho tem como base alguns sistemas que participaram no evento HAREM (Santos e Cardoso, 2007), organizado pela Linguateca, pelo que é necessário uma breve explicação do que é o HAREM e quais foram os seus objectivos.

O HAREM, segundo a Linguateca é "uma avaliação conjunta na área do reconhecimento de entidades mencionadas em português. Muito simplificadamente, é uma iniciativa que pretende avaliar o sucesso na identificação e consequente classificação automática dos nomes próprios na língua portuguesa".

Este evento teve duas avaliações, em que participaram vários sistemas de extracção de informação, para em conjunto, avaliarem os vários aspectos necessários para o reconhecimento de texto na língua portuguesa. O objectivo destes eventos foi entre outros, o de testar a capacidade de reconhecimento de EM em textos não estruturados. Esses sistemas tinham que identificar, classificar e anotar as EM encontradas.

A avaliação desses sistemas foi feita de acordo com regras pré estipuladas e com textos anotados manualmente, para que se pudesse, no fim dos testes comparar os resultados obtidos pelos vários sistemas com os textos anotados manualmente. Estes textos têm vários níveis de dificuldade, desde os textos para treino dos sistemas, colecção HAREM, até aos textos de avaliação, a colecção dourada (CD). No meu trabalho, para realizar as experiências que descrevo na secção 3.3 usei, para além de outros textos retirados aleatoriamente da Web, alguns dos textos da colecção HAREM.

Os sistemas participantes nos eventos HAREM usam modelos semânticos, nomeadamente ontologias, dicionários, expressões pré-definidas e outras formas que permitam agilizar o reconhecimento. Os sistemas criados no âmbitos destes eventos tinham quase todos o objectivo de reconhecer todas as entidades que fossem mencionadas num determinado texto mas haviam alguns que se preocupavam mais com

o reconhecimento de entidades geográficas, tentando reconhecer locais mencionados no texto e qual a relação existente entre eles.

Este trabalho é centrado exactamente no reconhecimento de entidades geográficas pelo que dos trabalhos relacionados, apresentados em seguida, alguns tiveram participação directa no 1º e no 2º HAREM.

2.2.2 O sistema REMBRANDT

O **REMBRANDT** (Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto) (http://xldb.di.fc.ul.pt/Rembrandt/) é um sistema que reconhece todas as entidades mencionadas (EM) em textos, na língua portuguesa e foi um dos sistemas participantes no 2º HAREM (REMBRANDT, 2008).

Este sistema usa a Wikipedia como principal base de conhecimento para a classificação de EM. De acordo com o autor, esta opção pelo uso da Wikipedia deve-se ao facto desta fornecer informação em formato XML ou SQL, o que permite uma consulta sobre um quase infinito repositório de informação permitindo dessa forma alargar a capacidade de pesquisa por mais informação acerca das EM reconhecidas.

Para o reconhecimento das EM este sistema passa por várias fases:

1ª - Recebe como entrada, texto simples, ficheiros HTML ou XML;

2ª - É feito um primeiro reconhecimento, usando um atomizador de entidades candidatas a EM através da leitura de cada frase do texto;

3ª - Usando uma interface específica, faz uma procura pelas EM identificadas na 1ª fase, separando em simultâneo as EM candidatas recorrendo para isso ao uso de regras gramaticais, fazendo deste modo uma desambiguação das EM. Esta interface é a ligação entre este sistema e a Wikipedia, pois faz um pré-processamento dos ficheiros fornecidos pela Wikipedia, classificando as possíveis EM e criando um relacionamento entre os resultados obtidos;

4ª - Caso haja um conflito entre EM reconhecidas, seja pela dificuldade da sua desambiguação seja por outro motivo como a falta de classificação de uma EM, o sistema usa o que chama de "*tribunal de EM"* onde são colocados argumentos entre a EM reconhecida pelo sistema na 1ª fase e alguma EM reconhecida na 2ª fase, como por exemplo, se aparece no texto o nome "Elias Monteiro", este pode não obter nenhuma classificação, na 2ª fase do REMBRANDT, no entanto pode ter sido classificado como sendo uma pessoa, na 1ª fase. Nestes casos o REMBANDT usa uma regra que atribui à entidade reconhecida a classificação mais comum, neste caso, pessoa.

5º- O reconhecimento de relações entre as EM é feito através do uso de heurísticas com base nas unidades, nas categorias e nas ligações entre páginas na Wikipedia. Estas heurísticas são aplicadas apenas a entidades não numéricas, como por exemplo, no caso de existir no texto dois nomes, *Elias Monteiro* e *António Elias Monteiro*, a heurística utilizada identificar que existe uma probabilidade de os dois nomes referirem-se à mesma pessoa.

2.2.3 O sistema CaGE

Este sistema tem como objectivo a atribuição de âmbitos geográficos (i.e. área geográfica referida num documento) a documentos textuais (Martins, 2008). O CaGE é um *sistema híbrido apoiado por dicionários e regras de desambiguação.* O funcionamento deste sistema baseia-se no uso de dicionários e de almanaques. Para efectuar o reconhecimento das EM, o CaGE passa por quatro etapas:

Etapa 1 - Identificação inicial das Entidades Mencionadas

Nesta fase, o sistema executa um processo de atomização do texto, recorrendo ao algoritmo baseado numa biblioteca de texto em Java. O algoritmo implementado por esta classe tem como base o implementado por (Gillam, 1999), que usa uma tabela contextual de pares de caracteres para separar as palavras, procurando os caracteres que ocorrem antes e depois de cada palavra. No caso especifico deste sistema, a tabela para a atomização é constituída por símbolos que indicam como são separadas as palavras consoante a pontuação. O resto da etapa processa-se da seguinte forma:

- As palavras identificadas no texto são guardadas como sequências, sendo que, o tamanho máximo dessa sequência é de seis palavras, se houver uma sequência maior, o sistema ignora-a;
- Depois de obter essa sequência de palavras, o CaGE vai filtrar as mesmas usando como critério o facto de começarem com uma letra maiúscula;
- Em seguida procura nos dicionários se existem as sequências filtradas, com a finalidade de mapear as mesmas com as entidades nos dicionários, registando a entidade mais geral no caso de haver algum conflito;
- O sistema usa expressões regulares para identificar entidades da categoria TEMPO que não se encontram definidas nos dicionários, como por exemplo, datas de calendário.

Etapa 2 - Classificação das entidades mencionadas e tratamento da ambiguidade

Nesta fase, tal como o título indica, este sistema vai classificar e desambiguar as entidades reconhecidas na primeira etapa procedendo da seguinte forma:

Primeiro, o sistema tenta resolver eventuais conflitos existentes entre entidades com mais do que uma classificação, usando regras definidas manualmente para encontrar a categoria e o tipo da entidade. Estas regras procuram por palavras-chave dentro do contexto do texto em que se encontra a entidade, procurando as duas palavras que ocorrem antes e depois da entidade em questão, por exemplo *Vendas Novas* terá duas identificações, *Vendas* e *Vendas Novas*. Neste caso o sistema opta pela designação mais generalizada que será *Vendas Novas*. Se persistirem, depois do passo anterior, entidades não desambiguadas, o sistema usa o algoritmo *round-robin classification* (Fürnkranz, 2002) fazendo assim uma classificação por escolha circular entre as várias categorias e tipos possíveis. De acordo com o autor, "*O argumento por detrás desta estratégia é o de que, escolhendo uma entidade diferente em cada situação ambígua e ir sequencialmente percorrendo o conjunto de atribuições possíveis, minimiza-se o número de erros introduzidos pelo sistema*".

Etapa 3 : Desambiguação completa de entidades geográficas e temporais

O processo de desambiguação das entidades geográficas e temporais processa-se da seguinte forma:

- Por cada entidade da categoria Local que foi identificada na etapa 2, o sistema procura por conceitos geográficos associáveis a essa entidade, usando o almanaque DIGMAP (Manguinhas et al., 2008).
- No caso de múltiplos resultados, estes são ordenados de acordo com a heurística "*um sentido por omisso*" (Martins et al., 2008), como por exemplo *Sintra*, está mais associada à *Vila de Sintra* do que ao *Concelho de Sintra*.

- No caso de ainda haver entidades não desambiguadas, o CaGE usa outra heurística, "*referentes relacionados por cada unidade de discurso*" (Martins et al., 2008) de forma a conseguir ordenar os conceitos geográficos subjacentes a essas entidades, como por exemplo no caso da frase "Lisboa fica no Distrito de Lisboa", o CaGE, neste caso faria uma ordenação das entidades de forma hierárquica, dando relevância a Lisboa como distrito.

  Quanto às entidades classificadas como sendo da categoria Tempo, estas passam pelo mesmo processo que as entidades da categoria Local.

Etapa 4: Atribuição de âmbitos geográficos e temporais aos documentos

Com base nas referências geográficas encontradas no texto, o sistema CaGE atribui um âmbito geográfico à totalidade do documento usando um algoritmo assente no uso de uma hierarquia de relações de inclusão entre os conceitos geográficos reconhecidos no texto (Amitay et al., 2004). Como fonte de dados para as relações é usado o almanaque DIGMPAP.

CaGE foi o único, dos trabalhos relacionados mencionados que participou nas duas edições do HAREM por isso importa fazer uma pequena comparação entre o objectivo da primeira participação e a segunda.

Na primeira participação, o CAGE, não sendo um sistema para reconhecimento de todas as entidades mencionadas de acordo com as regras do HAREM, focou-se no

reconhecimento de EM na categoria [Local] e conseguindo bons resultados nessa área. No segundo HAREM, o autor quis ir mais longe e acrescentou à categoria [Local] as categorias [Pessoa], [Organização] e [Tempo], tendo subdividido a categoria [Local] em subtipos, de modo a ter uma desambiguação das EM mais fidedigna, podendo, por exemplo reconhecer que o nome de uma pessoa tal como *Camilo Castelo Branco*, que pode ser confundido com o nome de uma localidade, no caso do exemplo, *Castelo Branco*, seja mais fácil de se identificar.

2.2.4 O sistema SEI-Geo

O SEI-Geo (Sistema de Extracção, Anotação e Integração de Conhecimento Geográfico) (SEI-Geo, 2008), participante do segundo HAREM foi o que conseguiu melhores resultados no reconhecimento de EM e no relacionamento entre estas. Faz parte de uma arquitectura de gestão de conhecimento geográfico designada por GKB. (*Geographic Knowledge Base),* (Chaves et al., 2005b).

O sistema foca-se essencialmente no reconhecimento de EM na categoria Local usando para isso duas geo-ontologias, a Geo-Net-PT01 que está limitada em 10 níveis hierárquicos, até à categoria freguesia e da WGO (*World Geographic Ontology*), que contém entidades administrativas e fisicas de todo o mundo, como países, rios e montanhas.

O SEI-Geo faz parte de um sistema de gestão do conhecimento (GKB) sendo um ambiente de extracção e integração de conhecimento geográfico contendo informação geográfica semi-estruturada. (Chaves, 2005) afirma que para a expansão do conhecimento contido no GKB são utilizados textos de onde se extrai informação. Esses textos são a entrada de informação para o SEI-Geo, que é o responsável por gerar uma representação estruturada do conhecimento geográfico extraído e integrá-lo no repositório do GKB.

Como se pode ver o SEI-Geo é apenas um módulo de um sistema maior, para este trabalho interessa-nos focar apenas neste módulo pelo que vou descrever o modo de funcionamento do mesmo aquando da sua participação no Segundo HAREM. O SEI-Geo é constituído por dois módulos, um para identificar candidatos a possíveis EM e

outro para classificar, extrair arbustos e anotar o conhecimento geográfico disponível em textos gerando uma representação numa forma estruturada.

Os módulos são constituídos por vários sub-módulos cujo funcionamento se descreve a seguir:

- **Identificador**

  O processo de reconhecimento das EM começa pela fase de *identificação*. Nesta fase, o sistema recebe como entrada, textos previamente segmentados em frases juntamente com um conjunto de padrões, conceitos e ocorrências de geo-ontologias. O resultado desta fase de identificação de potenciais candidatos a EM, são a entrada para a fase de *Classificação*;

- **Classificador**

  Este sub-módulo recebe o resultado do identificador e consulta as ontologias para reconhecer as EM e ver se existem relações entre as mesmas, marcando-as de forma a poderem, na fase seguinte serem estruturadas.

O segundo módulo é constituído pelos seguintes sub-módulos:

- **Extractor de arbustos**

  Depois de identificadas e classificadas as EM, este sub-módulo constrói o que é designado de arbustos, que são uma representação em árvore de entidades e das suas relações, como por exemplo se for mencionado "Av. 24 de Julho em Lisboa, que vai dar a Alcântara", o SEI-Geo vai construir uma representação gráfica com a seguinte representação da Fig. 3.

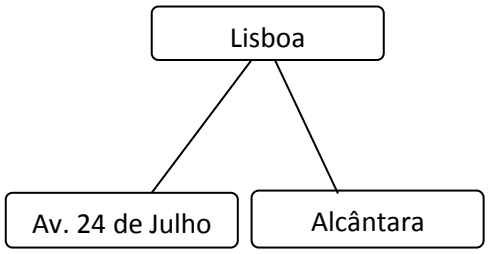


Fig. 1 – Exemplo de construção de um arbusto

Estes arbustos são constituídos por pelo menos duas entidades e representados através do uso do modelo de dados RDF (Ver Anexo A).

- **Anotador**

  Esta é a última fase deste módulo em que são anotadas as EM reconhecidas e as suas relações, sendo que essa anotação pode ser formatada como XML, HTML, JSON, etc.

2.2.5 O sistema SeRELeP

O SeRELep (Bruckschen et al., 2008), participante no Segundo HAREM, foca-se principalmente na identificação das relações entre as entidades, deixando que o pré-processamento dos textos seja feito por um outro sistema, de atomização de palavras, o PALAVRAS (Bick, 2000). As relações são identificadas com base em heurísticas que usam apenas as informações contidas no texto e as fornecidas pelo PALAVRAS. Este é um sistema que usa um dicionário apenas para relacionar as expressões encontradas no pré processamento com as categorias exigidas pelo HAREM, usando apenas regras linguísticas e o posicionamento das EM no texto, para identificar as relações existentes.

O SeRELep recebe como fonte de entrada de texto, ficheiros XML fornecidos pelo HAREM, fazendo a conversão do mesmo para texto plano, de modo a poder ser processado pelo PALAVRAS. Esta conversão é feita recorrendo ao que os autores chamam de SeRELep *tools*. Depois de tratado pelo PALAVRAS, o documento é dividido em três ficheiros, *token, pos e phrase*, representando níveis linguísticos, que são convertidos para o formato XCES (XML CES: *Corpus Encoding Standard for XML*, conforme http://www.xces.org/) recorrendo ao conversor Tiger2XCES (Bruckschen et al., 2008).

- **O ficheiro *Token***

  Neste ficheiro a informação é guardada de forma a saber a posição no texto do inicio e do fim de cada palavra.
- **ficheiro POS (*Part-Of-Speech*)**

Este é o ficheiro que guarda a informação de nível morfossintático, sendo referenciado, cada elemento do ficheiro Token, neste ficheiro.

- **O ficheiro *phrase***

  Este ficheiro descreve a informação a nível sintáctico, identificando frase, sujeitos, predicados e objectos, sendo estes grupos identificados como intervalos de elementos *token*.

Após estas duas etapas, o sistema vai então realizar o reconhecimento das relações, fazendo uso de heurísticas simples. Começando pela identificação da EM, o sistema tenta ver as relações que possam existir de inclusão (p.ex. dentro de, pertence a, etc) e de ocorrência (p.ex aconteceu em, ocorre em)

3. O sistema REMPT (Reconhecimento de Entidades Mencionadas em Português)

Criar um sistema que reconheça EM em textos não estruturados é uma tarefa relativamente simples (Linguateca, 2007), quando comparada com a tarefa proposta neste trabalho, pois esse tipo de sistemas não lida com as eventuais relações que possam existir entre as entidades, limitando-se a reconhecer as mesmas. Quando se fala em entidades geográficas, o problema sobe de complexidade pois a ambiguidade é de tal forma grande que podemos confundir facilmente o nome de um local com o nome próprio de uma pessoa. Por exemplo, na frase " Fui a Castelo Branco", Castelo Branco não oferece nenhuma confusão, mas se for lido dentro de uma frase como "Hoje li um livro de Castelo Branco", um sistema de reconhecimento de entidades geográficas pode facilmente ser confundido. Tal sistema pode inferir que existe uma entidade geográfica naquela frase, quando na verdade a entidade existe mas diz respeito a um nome. Estes e outros problemas foram sendo encontrados aquando da construção do meu sistema

Neste capítulo descrevo em detalhe o sistema REMPT para reconhecimento de entidades geográficas mencionadas em textos não estruturados, na língua portuguesa.

3.1 Requisitos e Arquitectura de Software

3.1.1 Requisitos

Requisito 1	**Entrada de texto não estruturado**
Descrição	
Deverá ser possível a entrada manual de texto não estruturado através da interface do REMPT	
Comentários:	
O inicio de processamento do texto para reconhecimento de EM requer que seja introduzido um texto não estruturado.	

Requisito 2	**Uso de Navegador Web**
Descrição	
O utilizador tem de usar a aplicação REMPT através de um navegador	
Comentários:	
Qualquer navegador que consiga efectuar pedidos e respostas usando o protocolo HTTP, pode ser usado para processar textos através do REMPT.	

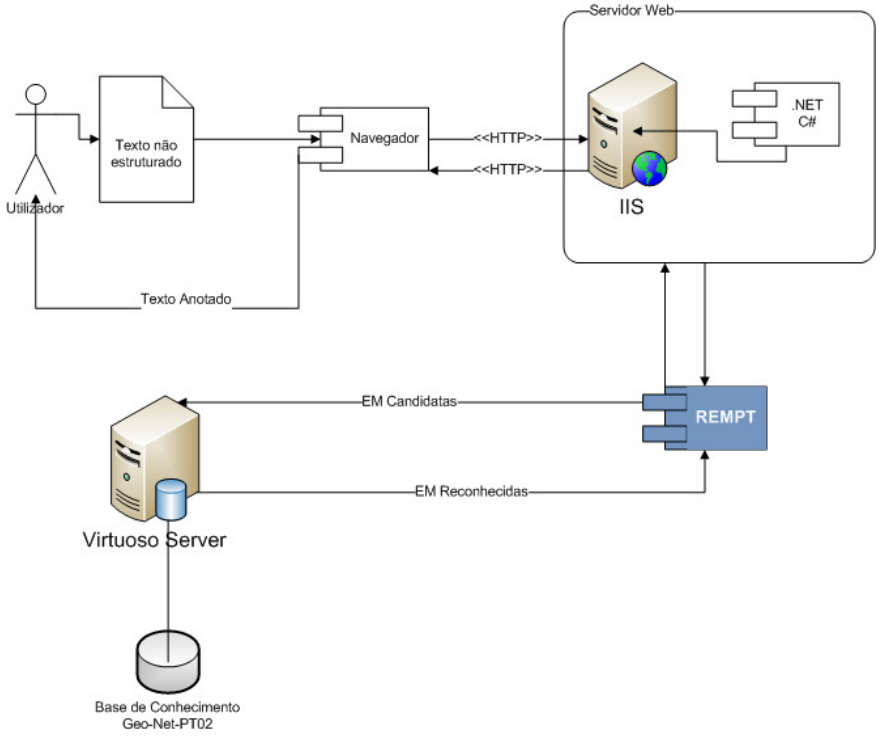
Requisito 3	**Servidor Internet IIS + Framework .NET 2.0 ou superior**
Descrição	
O REMPT necessita de ter instalado um servidor IIS (*Internet Information Services*) e a Framework .Net da Microsoft na sua versão 2.0 ou superior, sem os quais não será possível executar a aplicação.	
Comentários:	
A comunicação entre a aplicação e o utilizador é feita através de um navegador que por sua vez irá interagir com o IIS via HTTP	

Requisito 4	**Servidor de base de dados Virtuoso**
Descrição	
O REMPT necessita de ter instalado um servidor Virtuoso (http://sourceforge.net/projects/virtuoso/).	
Comentários:	
O servidor Virtuoso permite instalar a base de conhecimento Geo-Net-PT02, fornecida no formato RDF, criando um *endpoint* que permite efectuar consultas através de SPARQL.	

Requisito 5	**Acesso a base de conhecimento Geo-Net-PT02**
Descrição	
A base de conhecimento Geo-Net-PT02 é fornecida pelo Grupo XLDB, da Faculdade de Ciências da Universidade de Lisboa (http://www.linguateca.pt/geonetpt/geonetpt02/)	
Comentários:	
O servidor Virtuoso permite instalar a base de conhecimento Geo-Net-PT02, fornecida no formato RDF, criando um *endpoint* que permite efectuar consultas através de SPARQL .	

3.1.2 Arquitectura de software

A figura 2 apresenta a arquitectura de software necessária para o funcionamento do REMPT



A arquitectura de software do REMPT é descrita a seguir:

1. O REMPT recebe como dados de entrada textos não estruturados introduzidos manualmente pelo utilizador, através do uso de um navegador que use o protocolo HTTP, como é o caso do Internet Explorer, o Mozilla Firefox, o Chrome ou o Opera;
2. Depois de recebido esse texto é enviado via HTTP para o REMPT, onde é processado de acordo como descrito nas secções 3.1 a 3.3. ;
3. Após esse processamento, a lista de Entidades Mencionadas candidatas é verificada através do acesso à base de conhecimento instalada no servidor Virtuoso;
4. O servidor Virtuoso retorna uma lista com as entidades reconhecidas para o REMPT;
5. Depois de processado o texto, o REMPT devolve ao utilizador o texto anotado com as EM reconhecidas e as relações existentes.

3.2 Arquitectura do sistema REMPT

A Figura 2 mostra a arquitectura do REMPT, constituído por quatro etapas principais desde a entrada de texto pelo utilizador até à saída do mesmo texto, classificado, relacionado e anotado. Nas secções seguintes são descritas, em pormenor cada uma das etapas, explicando quais os problemas que existiram e como foram resolvidos.

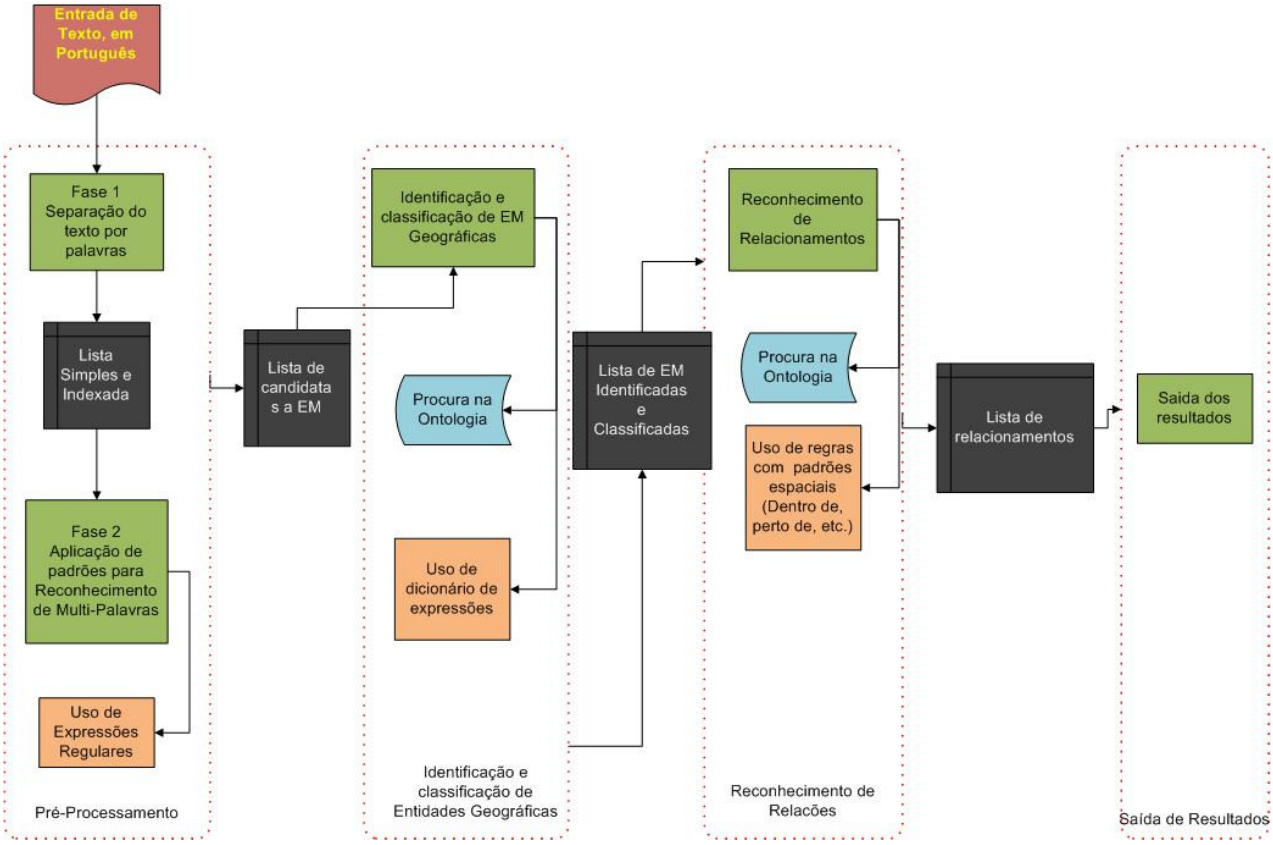


Fig. 3 - Arquitectura Geral do sistema REMPT.

3.2.1 Pré-Processamento

Após a entrada de texto pelo utilizador, o sistema inicia a etapa 1, o pré-processamento que tem como objectivo criar uma lista de possíveis candidatos a referências geográficas aliviando a carga de processamento para a fase posterior. O pré-Processamento é uma etapa constituída por duas fases, a da separação das palavras e a do tratamento dessas palavras separadas.

Fase 1: O pré processamento faz-se utilizando uma expressão regular que separa o texto pelas palavras que o constituem, retirando não só os espaços como também a pontuação. Esta fase é relativamente simples de executar e bastante rápida sendo as palavras guardadas numa lista simples e indexada.

Fase 2: Tomando como regra geral que o nome de um local começa por uma letra maiúscula ou um número, esta segunda fase do pré-processamento percorre a lista criada anteriormente procurando todas as palavras que comecem com uma letra maiúscula e guardando-as numa nova lista. Esta operação não é suficiente para garantir a identificação de toda a espécie de nomes possíveis para EM geográficas. Em seguida descrevo alguns dos principais problemas encontrados para efectuar esta segunda fase do pré-processamento e quais as soluções propostas e implementadas.

Problema: Reconhecimento de Entidades mencionadas multi-palavra.

Descrição: A lista criada na fase 1 apenas continha palavras que, no seu início tinham uma letra maiúscula. No entanto, um sistema de REM deve reconhecer entidades multi-palavra tais como *Macedo de Cavaleiros* ou *Vila Real de Santo António*, o que fez com que a lista anterior não fosse suficiente para garantir o reconhecimento desta forma de entidades. Como se pode observar na Tabela 2, (Chaves, 2009) fez um levantamento do número de Entidades, separadas pelo número de palavras que as constituem. Podemos observar que as entidades multi-palavra que são constituídas por mais de quatro palavras encontram-se mais concentradas nos nomes de localidades e de zonas, no entanto quando comparamos esses números com o número de entidades que têm até quatro palavras no seu nome, podemos concluir que o foco de reconhecimento deve ser nas entidades que contêm entre uma a quatro palavras.

Tipo	# de termos distintos	# de palavras nos termos					# de termos MP	Total ambiguidade	1 grama ambiguidade	
		1	2	3	4	∑>4			T	P
NUT1	3	1	0	0	2	0	2	3	0	0
NUT2	7	5	0	0	2	0	2	7	5	0
NUT3	30	8	11	8	3	0	22	6	2	4
região	2	0	1	0	1	0	2	0	-	-
província	11	**4**	6	0	1	0	7	5	2	1
distrito	18	15	2	1	0	0	3	18	15	0
concelho	323	203	27	68	22	3	121	301	193	1
ilha	11	0	1	6	4	0	11	1	-	-
freguesia	3.597	2.133	336	764	287	77	1.462	2.799	1884	51
localidade	26.924	10.851	4.098	9.661	**1.783**	**531**	16.073	3.655	2388	607
zona	3.593	1.201	540	1.233	**456**	**163**	2.392	1.241	804	55
Total	34.519	14.421	5.022	11.741	2.561	774	-	-	-	-

Tabela 2- Descrição quantitativa da Geo-Net-PT01 (Chaves, 2009).

Solução: A solução encontrada para este problema foi o uso de artigos definidos e preposições (ex: o, a, as, os, de, do, da, dos, das), recorrendo a uma expressão regular e percorrendo a lista palavra a palavra até encontrar esses artigos e identificar este tipo de entidade.

Variáveis em uso:

- palavraCorrente - Representa a palavra encontrada pela expressão regular;
- palavraSeguinte – Representa a palavra que vem imediatamente a seguir à palavra corrente;
- palavraAnterior – Representa a palavra imediatamente anterior à palavraCorrente;
- palavras [] – Representa todas as palavras do texto
- contador – Representa a posição onde foi encontrado o artigo, no texto
- candidata [ ] – Representa, em caso de identificação positiva a entidade mencionada, após a conclusão do algoritmo.

Outras definições

- frase [ ]– No contexto do algoritmo entende-se como frase a linha onde foi encontrada a palavraCorrente.

Quando o REMPT detecta um artigo executa o algoritmo representado na Fig. 5.

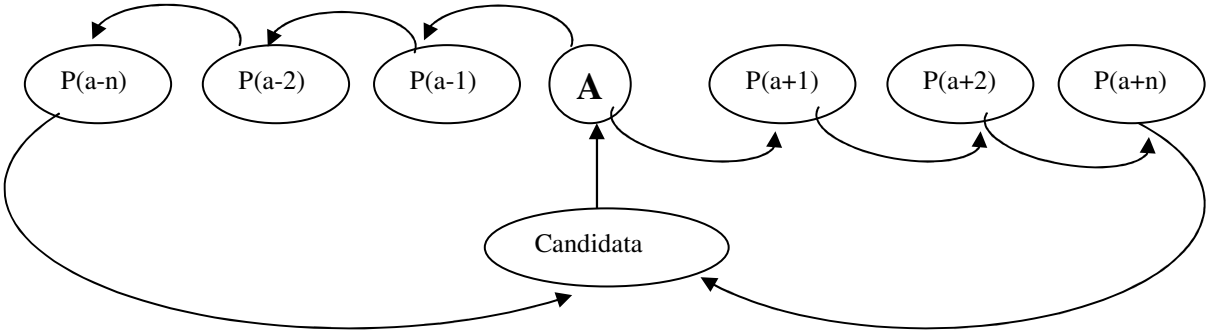


Fig. 4 – Algoritmo para encontrar entidades multi-palavra

Quando encontra um artigo (A), o algoritmo executa um ciclo verificando se as palavras anteriores (P(a-n)) a esse artigo iniciam-se por letra maiúscula. Em caso positivo, cada palavra é concatenada na variável candidata. Este ciclo é efectuado até que seja encontrada uma palavra que não se inicia com letra maiúscula. Quando isso acontece são verificadas as palavras posteriores ao artigo (A) seguindo as mesmas regras. Quando termina este último ciclo, a entidade guardada na variável candidata é adicionada à lista das entidades candidatas.

Pseudo-código para identificação de entidades multi-palavra:

```
palavraCorrente=artigo
Se contador >= 1 Então
      Enquanto contador>=0 faz
            palavraAnterior=palavras [contador -1]
            Se  palavraAnterior  começar  com  letra  maiúscula
      Então
                  candidata   =   palavraAnterior   +   "   "   +
            candidata
            Senão
                  candidata = candidata+palavraCorrente;
                  Sai do ciclo Enquanto
            Fim do Se
      Contador= contador-1
      Fim de Enquanto
      candidata = candidata + palavraCorrente
      Enquanto contador <totalPalavras faz
            palavraSeguinte= palavras[contador+1]
            Se  palavraSeguinte  começar  com  letra  maiuscula
      Então
                  candidata =candidata+ " " + palavraSeguinte
            Fim de Se
      Fim de Enquanto
Fim de Se
```

Pese embora se saiba que existem entidades que não serão reconhecidas, por terem nomes com mais de quatro palavras ou por conterem mais de um artigo na mesma entidade, este algoritmo resolve a maior parte dos problemas existentes com as entidades multi-palavra.

Problema: Reconhecimento de EM multi-palavras, separadas por hífen (-).

Descrição: Apesar de no problema anterior ter resolvido a questão das localidades com mais do que uma palavra no seu nome, as entidades reconhecidas com aquela solução são apenas as que têm um espaço imediatamente antes do artigo encontrado. O outro caso que acontece nos nomes dos locais, tem a ver com a separação das várias palavras de uma localidade recorrendo ao uso do hífen, como por exemplo, *Linda-a-Velha.*

Solução: Para contornar este problema, usei a mesma técnica da solução anterior, começando pela procura do artigo que separa as palavras como em Linda-a-Velha, usando a mesma expressão regular.

A diferença nesta verificação prende-se com o facto de que na pesquisa anterior, verificava se antes de cada artigo encontrado existia um espaço, neste algoritmo começo por ver se o carácter imediatamente anterior à localização do artigo é um hífen. Em caso afirmativo, o REMPT procede da seguinte forma:

```
Se carácter anterior à palavraCorrente= hífen e carácter posterior
à palavraCorrente= hífen então
      Procura  palavraAnterior à palavraCorrente
      Se palavraAnterior começar com letra maiúscula então
            candidata = palavraAnterior + "-" + palavraCorrente
            Procura  palavraSeguinte à palavraCorrente
            Se palavra seguinte começar com letra maiúscula então
                  candidata = candidata+ "-" + palavraSeguinte
      Se palavraAnterior não começa com letra maiúscula então
Procura números
```

Apesar desta solução identificar entidades como *Linda-a-Velha*, não conseguia identificar outras entidades como por exemplo *Agualva-Cacém*. Por isso apliquei outra regra, que é usada quando na posição corrente da lista não se encontra um artigo. Nesse caso verifico se existe um hífen imediatamente antes da palavra corrente. Em caso afirmativo, o REMPT age do seguinte modo:

```
Se carácter anterior à palavraCorrente é hífen e carácter
posterior à palavraCorrente não é hífen então
      Procura palavraAnterior à palavraCorrente
            Se palavraAnterior começar com letra maiúscula então
                  candidata = palavraAnterior+"-"+palavraCorrente
```

Problema: Reconhecimento de EM que começam por números.

Descrição: Apesar de não ser nesta fase que é feita a classificação por tipo, das EM, é nesta fase que guardo as candidatas a entidades geográficas, e por exemplo a expressão "1º de Maio" pode corresponder a uma Rua, Travessa, Avenida, etc.

Solução: Para resolver este problema tomei como princípio que a seguir a um número vem um artigo definido ou uma preposição, como em "1º **de** Maio". Após esta decisão, usei duas expressões regulares, uma para verificar as ocorrências de números com os

símbolos º e ª e outra para a verificação da existência de dígitos, tal como em "25 de Abril". Esta verificação faz-se quando se detecta que a palavra anterior não começa com uma letra maiúscula, procedendo-se da seguinte forma:

```
      Se palavraAnterior não começar com letra maiúscula então
            Se palavraAnterior é um número com º ou ª então
                  candidata = palavraAnterior+ " " + palavraCorrente
                  Procura palavraSeguinte
                  Se palavraSeguinte começa com letra maiúscula então
                  candidata = candidata+ palavraSeguinte

                  Senão
                        Se palavraAnterior é um número inteiro
positivo·               então
                        candidata = palavraAnterior+" "+
palavraCorrente
                        Procura palavraSeguinte
                  Se palavraSeguinte começa com letra maiúscula então
                  candidata = candidata+ palavraSeguinte
```

No fim desta fase é construída uma nova lista com os possíveis candidatos a EM geográficas que vai ser usada em conjunto com a Geo-ontologia, para identificar e classificar as EM geográficas correctas.

3.2.2 Identificação e Classificação de Entidades Geográficas

Depois de identificadas as possíveis candidatas a entidades geográficas, nesta etapa o sistema percorre a lista criada na 2ª fase do pré-processamento e vai identificar realmente, daquela lista, o (s) elemento (s) que corresponde (m) a entidades geográficas de modo a que na 3ª etapa do processamento possam ser usadas apenas as entidades reconhecidas, para proceder ao reconhecimento de relações entre as várias entidades, caso existam.

Dos sistemas que serviram de base para a elaboração deste trabalho observei que (Delboni, 2005) opta por fazer uma pesquisa prévia de informação sobre a área geográfica a estudar, carregando depois os dados numa base de dados. O REMBRANDT opta por carregar para uma base de dados os dados fornecidos pela Wikipedia e outros optam pelo uso de uma geo-ontologia (Chaves, 2009). Para obter uma identificação de entidades geográficas o mais correctamente possível, recorri ao

uso da geo-ontologia Geo-Net- PT02 (Lopez-Pellicer et al., 2009), criada pelo Grupo XLDB, da Faculdade de Ciências da Universidade de Lisboa pois esta geo-ontologia já contém informação geográfica sobre Portugal.

Esta geo-ontologia está organizada de forma a facilitar o acesso à informação que necessitamos a partir de um nome ou de um tipo de entidade (Cidade, Concelho, Rua, etc.), como exemplo temos o nome de um local "Seixal" que de acordo com (Santos et al, 2008) existem 42 locais com o mesmo nome em Portugal, tal como este existem outros nomes que, pela sua ambiguidade, podem pertencer a vários tipos cabendo ao sistema tentar desfazer as dúvidas quanto à sua classificação.

De acordo com a página web relativa a esta geo-ontologia, http://xldb.di.fc.ul.pt/wiki/Geo-Net-PT_02, a mesma contém todos os dados geográficos administrativos de Portugal, incluindo sítios da Web portuguesa e os seus âmbitos geográficos, como é o caso do exemplo dado nessa página relativamente ao site da câmara municipal de Lisboa, http://www.cm-lisboa.pt, que tem como âmbito geográfico o concelho de Lisboa.

A geo-ontologia Geo-Net-PT2 é apresentada num modelo de dados RDF, um modelo simples que usa uma sintaxe baseada em XML permitindo dessa forma uma interacção com os dados mais facilitada.

A estrutura de cada elemento da geo-ontologia, segue a norma de construção de elementos para o modelo de dados RDF (Ver Anexo A), colocando elementos dentro de um elemento padrão, <rdf:Description rdf:about="#sintra-AF240249"), que no caso desta geo-ontologia, vai-nos servir para identificação de todas as entidades geográficas presentes na geo-ontologia, que têm uma ou mais relações com Sintra. A geo-ontologia contém outros elementos, dentro do elemento *description*, como por exemplo o elemento <gn:type rdf:resource="#localidade-ATLOC"/> que determina o tipo de entidade a que nos referimos, neste caso, uma localidade, poderíamos estar a referir-nos à Vila de Sintra, que apesar de ser a mesma localidade, é uma especificação do local. Existem outros elementos que permitem a identificação de arruamentos, transportes e localidades adjacentes, entre outros.

No exemplo abaixo, podemos observar uma parte do resultado de uma procura por Sintra, embora este exemplo esteja bastante reduzido, pois Sintra tem muito mais informação associada. Neste exemplo podemos ver que Sintra é uma localidade (*type*) e contém arruamentos (*hasPar*t), como por exemplo a Rua General Firmino Miguel.

```
<rdf:Description rdf:about="#sintra-AF240249"> -Nome da Entidade
   <rdf:type
rdf:resource="http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing"/>
   <rdf:type      rdf:resource="http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-
net#Feature"/>
   <dcterms:title   xml:lang="pt">Sintra</dcterms:title>  -  Título  para  esta
localidade
   <rdfs:label xml:lang="pt">Sintra (Localidade)</rdfs:label>
   <gn:inDomain rdf:resource="#GeoAdministrative"/>
   <gn:type rdf:resource="#localidade-ATLOC"/> - Tipo de Entidade
   <gn:lineage rdf:resource="#SRC-CTT"/>-Origem dos dados, neste caso, os
CTT
   <gnpt:preferred rdf:resource="#sintra-pt"/- Nome utilizado por omisão
   <gnpt:hasPart  rdf:resource="#general_firmino_miguel-AF240274"/>  -Faz
parte de
   <gnpt:hasPart rdf:resource="#2710-567-AF240273"/>
   <gnpt:hasPart rdf:resource="#garrett-AF240272"/>
   <gnpt:hasPart rdf:resource="#doutor_miguel_bombarda-AF240270"/>
 </rdf:Description>
```

No sistema CaGE (Martins, 2008) foram identificadas expressões que estão normalmente associadas a referências geográficas (Tabela 3). Para a identificação e classificação das EM, o REMPT usa nesta etapa um dicionário de expressões que conjuga as expressões designadas como *identificadores*, da tabela do sistema CaGE, adicionando outras expressões e suas abreviaturas, quando existem, representando os tipos de arruamentos existentes em Portugal.

Tipo de expressão	Expressão
Identificadores	cidade, município, distrito, rua, avenida, rio, ilha, montanha, vale, país, continente, zona, região, condado, freguesia, deserto, província, povoado, aldeia, monte, vila, república, península
Localização	fora de, nos arredores de, dentro de, entre, em acima, ao longo, atrás, acima, ao lado, à esquerda, à direita
Distância Relativa	adjacente, longe de, perto de, próximo de
Orientação	este, norte, sul, oeste, oriente, ocidente, sudeste, sudoeste, nordeste, noroeste
Outras	"cidades como", "e outras cidades", "cidades, incluindo", Expressões "cidades, especialmente", "uma das cidades", "cidades tais como", padrões semelhantes para outros identificadores

Tabela 3 -Expressões de contexto associadas a referências geográficas (Martins, 2007).

Expressões que indicam tipos de Arruamento em Portugal	R., Praceta, Prct., Av., Alameda, Praça, Beco, Estrada, Estr. Travessa, Largo, Calçada

Tabela 4 - Expressões adicionadas.

O uso deste dicionário serve para separar as EM, de modo a que fique claro quais as EM que se referem apenas a locais, sem classificação nenhuma em especial e as que têm que ter uma classificação associada ao tipo de local. Isto permite ainda que a classificação seja feita o mais correctamente possível evitando consultas desnecessárias à geo-ontologia.

O modo de funcionamento deste processo é o seguinte:

```
Procura no texto por cada uma as expressões do dicionário
Se encontrou uma das expressões então

Procura palavraSeguinte no texto
      palavraActual=palavraSeguinte
      Procura palavraActual na lista de candidatas
      Se encontrou e a posição marcada corresponde à posição da
palavraCorrente então
      Candidata= expressão encontrada + palavra Actual
```

Depois de guardada a associação de tipo à entidade identificada, como por exemplo a *Rua* 1º de Maio, o REMPT procede à procura, na geo-ontologia por essa entidade. Caso seja confirmado que é uma entidade constante da geo-ontologia, a entidade é guardada noutra lista, a ser usada no reconhecimento de relacões. Caso contrário, a entidade é descartada, na forma como foi concatenada no processo, assumindo-se que não se trata de uma entidade geográfica. No entanto, a palavra existente na lista da 2ª fase mantém-se pois, sozinha pode ter um significado geográfico.

Para todas as EM, candidatas na lista da 2ª fase, mas que não ficaram associadas a nenhuma expressão, o REMPT procura na geo-ontologia por correspondências. No caso de serem encontradas, essas EM são marcadas como sendo realmente EM e por isso, transportadas para a lista que irá transitar na etapa 3.

Tal como na 1ª etapa, descrevo alguns dos problemas encontrados:

Problema: EM multi-palavras demasiado longas.

Apesar do REMPT conseguir identificar na 2ª fase do pré-processamento, EM como *Vila Real de Santo António*, verifiquei que quando tenta classificar uma rua como por exemplo, a *Rua da Cintura do Porto de Lisboa*, falha, pois reconhece esta entidade como sendo "Porto de Lisboa" o que não corresponde à verdade.

Solução: A solução para este problema passa pela optimização do algoritmo de reconhecimento de EM multi-palavra de modo a poder reconhecer mais do que um artigo dentro da mesma sequência de palavras. Tal como referido em 3.1, o número de casos destes é bastante reduzido, comparativamente com os restantes.

Problema EM sem tipo geográfico, com mais do que uma referência

Este tipo de problema é bastante comum, por exemplo, se realizarmos uma pesquisa por Lisboa, a geo-ontologia irá retornar várias referências, a Lisboa Distrito, Lisboa Concelho ou Lisboa Localidade.

Solução: Quando existe este tipo de conflito optei, tal como foi feito noutros sistemas referenciados neste trabalho pelo uso da heurística, "*um sentido por omisso*" (Martins et al, 2008), que diz que devemos optar pelo conceito mais lato da entidade geográfica em causa. No caso do exemplo acima, Lisboa seria classificada como Lisboa, Concelho, pois é a forma mais conhecida de nos referirmos a Lisboa.

Optei também por outra heurística, "*referentes relacionados por cada unidade de discurso*" (Martins e tal, 2008), que tem como princípio a identificação das EM recorrendo à hierarquia dos tipos geográficos. Esta heurística é usada apenas para as entidades que não consigam ser identificadas usando as outras heurísticas, como por exemplo na frase "A Cidade de Lisboa fica no Distrito de Lisboa", o sistema iria atribuir como classificação a Lisboa, o tipo Distrito.

3.2.3 Reconhecimento de Relações

Nesta 3ª etapa, o REMPT recebe como entrada de dados, a lista, já classificada das entidades reconhecidas na etapa 2. O reconhecimento de uma relação entre entidades geográficas nem sempre é fácil pois existem na língua portuguesa várias formas de nos referirmos ao mesmo lugar. Por exemplo, se dissermos, "*Vou a Lisboa que fica em Portugal*", podemos depreender facilmente que *Lisboa faz parte* de *Portugal,* no entanto se dissermos o mesmo desta forma "*Vou a Lisboa, Portugal*" o REMPT não será capaz de identificar de imediato a relação existente entre Lisboa e Portugal. Nestas situações o sistema irá recorrer à geo-ontologia para procurar as duas entidades identificadas e classificadas, *Lisboa*, como *cidade* e *Portugal* como *país*. Ao encontrar *Portugal* irá encontrar uma relação de inclusão com *Lisboa*, ou seja ficará uma relação *Portugal*-***Inclui***-*Lisboa*.

O problema do reconhecimento de EM é um dos maiores que se põe quando falamos em sistemas de reconhecimento de entidades geográficas. Para este trabalho resolvi usar a abordagem utilizada no sistema CaGE, através do uso das expressões

identificadas na Tabela 3, nas categorias de *Localização, Distância Relativa, Orientação* e *outras* aplicadas a expressões regulares, que procuram no texto ocorrências para essas expressões e relacionam as EM que estiverem no contexto da frase, parágrafo ou texto. Depois de aplicada esta abordagem, se persistirem EM sem relação entre elas, recorro a outro método, o hierárquico, sempre que possível.

O método hierárquico consiste em procurar cada uma das EM restantes, directamente na geo-ontologia, pesquisando nos tipos de relacionamento existentes, *isAdjacentTo, isPartOf, hasPartOf* por cada uma das outras EM, se encontrar alguma, é-lhe atribuído esse tipo de relacionamento. Por exemplo, na frase "A loja fica em Lisboa, Av. de Berna", a relação explícita para o ser humano é a de que a loja fica em Lisboa, no entanto não há uma indicação de onde fica a Av. de Berna, nem sequer se fica em Lisboa. Nestes casos, o REMPT reconhece "Lisboa" e "Av. de Berna" como EM, procura na geo-ontologia por *Lisboa*, concelho (e não distrito, uma vez que a seguir tem uma EM identificada como um arruamento) e com o resultado obtido procura por *Av. de Berna* o que resulta numa relação de inclusão da *Av. de Berna,* em *Lisboa*.

Problemas e Soluções

Por exemplo, na frase "Hoje fui ao Freeport em Alcochete que fica perto de Lisboa", o REMPT vai identificar *Alcochete* e *Lisboa* como sendo duas EM geográficas, depois, como encontra a expressão "perto de" vai criar uma relação de proximidade entre as duas entidades.

No exemplo acima, foi muito fácil identificar o tipo de relação existente entre as duas EM, no entanto, no decorrer dos testes deparei-me com alguns problemas dos quais destaco a falta de indicadores de relações. Este tipo de problema ocorre quando não temos nenhuma pista, dentro da frase que nos permite relacionar duas ou mais entidades, entre si. Por exemplo, se tivermos a frase, "Hoje passeei por Lisboa, Alcochete, Montijo e Barreiro" não está implícito qual a relação que existe entre estas entidades, logo o REMPT não encontrará uma forma de classificar estas relações.

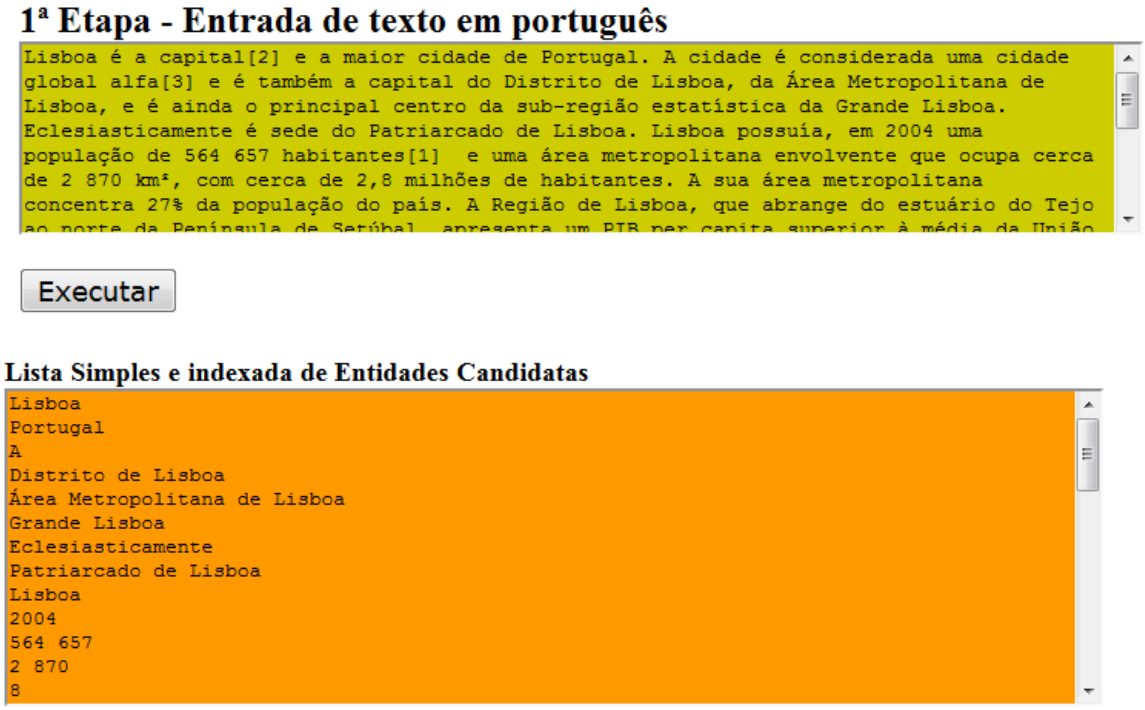
Quando existe um problema destes, o REMPT vai procurar na geo-ontologia por cada uma das entidades identificadas e saber se existe uma ou mais relações entre elas.

Por exemplo, na geo-ontologia, se procurarmos por *Alcochete* conseguimos saber que o *Montijo* tem uma relação de proximidade (*isAdjacentTo*) e também conseguimos saber que Alcochete pertence ao Distrito de Setúbal (isPartOf). Portanto podemos procurar as outras entidades dentro do distrito. Neste caso, o REMPT será capaz de reconhecer as relações entre Alcochete, Montijo e Barreiro pois todos pertencem ao distrito de Setúbal, no entanto não conseguirá estabelecer uma relação com Lisboa.

3.2.4 Saída de Resultados

Nesta ultima etapa do REMPT, as entidades reconhecidas vão ser anotadas e destacadas para que se possa ver, no texto, o que foi anotado e as suas relações.

A figura 4 apresenta as várias fases de processamento do texto não estruturado, desde a sua entrada até ao resultado final.



2ª Etapa -Identificação e Classificação de Entidades Geográficas

Candidata	idDirecta	idOntologia
Lisboa	Lisboa - Sem Classificação	Lisboa (Concelho)
Portugal	Portugal - Sem Classificação	Portugal (Pais)
A	A - Sem Classificação	Sem Classificação
Distrito de Lisboa	Distrito de Lisboa - Distrito	Lisboa (Distrito)
Área Metropolitana de Lisboa	Área Metropolitana de Lisboa - Sem Classificação	Sem Classificação
Grande Lisboa	Grande Lisboa - Sem Classificação	Grande Lisboa (NUT3)
Eclesiasticamente	Eclesiasticamente - Sem Classificação	Sem Classificação
Patriarcado de Lisboa	Patriarcado de Lisboa - Sem Classificação	Sem Classificação
Lisboa	Lisboa - Sem Classificação	Lisboa (Concelho)
2004	2004 - Sem Classificação	Sem Classificação
564 657	564 657 - Sem Classificação	Sem Classificação

3ª Etapa- Reconhecimento de Relações

EM	Tipo de Relação	Com EM
Lisboa (Concelho)	Faz Parte de	Lisboa (Distrito)
Lisboa (Concelho)	Faz Parte de	Grande Lisboa (NUT3)
Lisboa (Distrito)	Faz Parte de	Lisboa (Distrito)
Lisboa (Distrito)	Faz Parte de	Grande Lisboa (NUT3)
Lisboa (Concelho)	Faz Parte de	Lisboa (Distrito)
Lisboa (Concelho)	Faz Parte de	Grande Lisboa (NUT3)
Lisboa (Concelho)	Faz Parte de	Lisboa (Distrito)
Lisboa (Concelho)	Faz Parte de	Grande Lisboa (NUT3)
Odivelas (Concelho)	Faz Parte de	Lisboa (Distrito)
Odivelas (Concelho)	Faz Parte de	Grande Lisboa (NUT3)
Oeiras (Concelho)	Faz Parte de	Lisboa (Distrito)
Oeiras (Concelho)	Faz Parte de	Grande Lisboa (NUT3)
Amadora (Concelho)	Faz Parte de	Lisboa (Distrito)
Amadora (Concelho)	Faz Parte de	Grande Lisboa (NUT3)
Lisboa (Concelho)	Faz Parte de	Lisboa (Distrito)
Lisboa (Concelho)	Faz Parte de	Grande Lisboa (NUT3)
Seixal (Concelho)	Faz Parte de	Peninsula de Setúbal (NUT3)
Barreiro (Concelho)	Faz Parte de	Peninsula de Setúbal (NUT3)
Moita (Concelho)	Faz Parte de	Peninsula de Setúbal (NUT3)
Montijo (Concelho)	Faz Parte de	Peninsula de Setúbal (NUT3)
Grande Lisboa (NUT3)	Inclui	Lisboa (Concelho)
Grande Lisboa (NUT3)	Inclui	Lisboa (Concelho)

4ª Etapa - Saída de Resultados

Lisboa é a capital[2] e a maior cidade de Portugal. A cidade é considerada uma cidade global alfa[3] e é também a capital do Distrito de Lisboa, da Área Metropolitana de Lisboa, e é ainda o principal centro da sub-região estatística da Grande Lisboa. Eclesiasticamente é sede do Patriarcado de Lisboa. Lisboa possuía, em 2004 uma população de 564 657 habitantes[1] e uma área metropolitana envolvente que ocupa cerca de 2 870 km², com cerca de 2,8 milhões de habitantes. A sua área metropolitana concentra 27% da população do país. A Região de Lisboa, que abrange do estuário do Tejo ao norte da Península de Setúbal, apresenta um PIB per capita superior à média da União Europeia, que faz desta a região a mais rica de Portugal. O concelho de Lisboa tem 83,84 km² de área, e apresenta uma densidade demográfica de 6 734,94 hab./km². O concelho subdivide-se em 53 freguesias, encontrando-se em estudo a formação de uma nova freguesia, que abrangeria a zona do Parque das Nações. Faz fronteira a norte com os municípios de Odivelas e Loures, a oeste com Oeiras, a noroeste com a Amadora e a sudeste com o estuário do Tejo. Por este estuário, Lisboa une-se aos concelhos da Margem Sul: Almada, Seixal, Barreiro, Moita, Montijo e Alcochete.

Fig 5. Um exemplo de texto processado pelo REMPT

3.3 Experiências

3.3.1 Experiência 1

Para esta experiência foram usados textos retirados da Wikipedia relacionados com distritos ou localidades específicas, como por exemplo Lisboa, Vila Real de Santo António, Freixo de Espada à Cinta e outras. Este tipo de experiência permitiu aferir o grau de eficácia e abrangência do REMPT. Foram usados 50 textos, contendo 200 EM geográficas

Modo de avaliação:

Para avaliar a quantidade de EM geográficas reconhecidas pelo REMPT, os textos escolhidos foram marcados por mim, manualmente, assinalando as EM geográficas e as suas relações. Depois desta marcação foi executado o REMPT e foi feita uma comparação entre o que foi marcado manualmente e o resultado obtido. Os textos foram retirados da lista constante desta página http://pt.wikipedia.org/wiki/Categoria:

Cidades_de_Portugal.

Resultado da avaliação

Em termos de Abrangência:

Das 200 EM marcadas manualmente foram identificadas correctamente 180.

Em termos de Eficácia:

90% das EM foram identificadas e 70% das relações entre EM foram reconhecidas correctamente. As que não foram reconhecidas, na sua maioria foi devido aos problemas descritos na secção 3.2.2.

3.3.2 Experiência 2

Para esta experiência foram usados textos da colecção do Segundo HAREM, disponíveis em http://www.linguateca.pt/aval_conjunta/HAREM/

colSegundoHAREM.xml.

Embora os textos referidos estejam preparados para os sistemas que reconhecem todo o tipo de entidades mencionadas, consegui realizar alguns testes de modo a compreender a capacidade de reconhecimento de EM geográficas e suas relações.

Considerações sobre a experiência 2

O REMPT conseguiu identificar relações que continham as expressões utilizadas como léxico e as que se enquadrando no problema "Inexistência de identificador de relação" (Ver secção 3.2.3), conseguiram, através da geo-ontologia ser identificadas.

No reconhecimento das entidades, houve algumas entidades, como por exemplo cidades estrangeiras, que foram marcadas como EM, no entanto, como não têm nenhuma classificação existente na geo-ontologia, não é marcada nenhuma relação entre elas excepto se essa relação estiver explícita no texto.

Por exemplo, no texto "BUDAPESTE RECEBE, amanhã e terça-feira, os líderes dos 52 países membros da Conferência de Segurança e Cooperação na Europa (CSCE), para uma cimeira onde a Rússia vai tentar limitar a influência crescente da NATO no Leste europeu e ao mesmo tempo reforçar o seu papel nas antigas repúblicas soviéticas." As entidades reconhecidas seriam, Budapeste, Europa, Rússia mas o REMPT não conseguiria reconhecer uma relação entre estas entidades.

Devido à quase inexistência de EM portuguesas neste texto não é possível aferir os resultados em termos de eficácia e abrangência.

3.4 REMPT e os Trabalhos Relacionados

Nesta secção explico quais as semelhanças entre o REMPT e os trabalhos descritos na secção 2.2.

- **REMBRANDT**

  Apesar do REMBRANDT usar como base artigos fornecidos pela Wikipedia, que são convertidos para uma base de dados, periodicamente, O REMPT não usa este tipo de abordagem, pois o REBRANDT reconhece todas as EM, enquanto que ao REMPT apenas me interessa reconhecer as EM geográficas. No entanto, tal como neste sistema o REMPT faz um pré-processamento dos textos para o reconhecimento das candidatas a EM fazendo o uso de algumas regras gramaticais básicas que permitem o reconhecimento de potenciais EM.

- **CaGE**

  Apesar do objectivo deste sistema ser quase idêntico ao do REMPT, as semelhanças ficam-se pela atomização das frases, embora o REMPT faça de maneira diferente, o processo produz quase o mesmo resultado. O resto do processo de identificação, classificação e relacionamentos diferencia-se deste sistema, no número de dicionários e almanaques que uso.

- **SEI-Geo**

  Este sistema é o que mais semelhança tem com o REMPT pois baseia-se também no uso de geo-ontologias e no reconhecimento de padrões na 1ª fase. No entanto, o REMPT usa apenas a geo-ontologia Geo-Net-PT02 completa.

- **SEReLEP**

  Este sistema recorre a regras sintácticas, morfológicas e gramaticais para reconhecer relações do tipo, *identidade*, *inclusão* e *ocorre_em*, sem recurso

a nenhuma geo-ontologia ou outra espécie de almanaque. Apesar de o REMPT também reconhecer os mesmo tipos de relacionamento, fá-lo usando uma geo-ontologia e os tipos de relacionamento contidos nela, como por exemplo "Adjacente" (*isAdjacent*) ou "pertence a (*isPartOf*)".

4. Considerações finais

4.1 Conclusões

Neste trabalho descrevi sistemas implementados para o português para o reconhecimento de EM em textos, alguns, apenas direccionados para o reconhecimento de entidades geográficas. Neste trabalho também implementei um sistema de reconhecimento de EM em textos não estruturados em português, o REMPT.

O REMPT consegue extrair dos textos, as entidades geográficas, utilizando expressões regulares, tanto para a separação como mais tarde para encontrar relações entre elas. Esta última tarefa, as relações revelou-se a mais difícil de todas, pois existem várias formas, em português, de nos referirmos às mesmas coisas. Por exemplo, "Lisboa" pode ser referenciada como "LX", portanto se tivermos uma frase como "Hoje fui a LX na margem norte do rio Tejo", a relação de proximidade entre o Rio Tejo e Lisboa não será reconhecida. No entanto, se houver identificadores explícitos de distância, proximidade ou outros, o REMPT consegue encontrar as relações entre as entidades. A parte de classificação das entidades em EM. foi facilitada graças ao uso da geo-ontologia Geo-Net-PT02 que contém a informação necessária para sabermos identificar e classificar as entidades.

O REMPT pode servir de base para a criação de um sistema inteligente que possa juntar um motor de recuperação de informação com um motor de extracção de informação, trazendo para o domínio da Web, em português o conceito de Web Semântica associado à língua portuguesa. Assim, esses sistemas poderão responder a perguntas como "Quais são os pontos de interesse perto de Arraiolos?" com a resposta estruturada de modo a que o utilizador tenha de imediato acesso a essa informação. O uso de várias ontologias, juntando vários domínios de conhecimento já começa também a ser uma realidade, por isso, para além das entidades geográficas, será possível relacionar o conhecimento retirado da Geo-Net-PT02 e relacioná-lo com outras geo-ontologias, como por exemplo a Geonames.

Embora ainda existam limitações, o desenvolvimento deste tipo de software para a língua portuguesa tem vindo a evoluir. Com as avaliações efectuadas pelos eventos HAREM será possível avaliar esta evolução de modo a podermos ter no futuro este tipo de motores de busca, inteligentes e que não "entopem" os utilizadores com informação desnecessária.

4.2 Limitações

Para a realização deste trabalho foram muitas as limitações encontradas. A seguir descrevo as que considero mais pertinentes.

- Falta de informação detalhada de como as tecnologias são realmente utilizadas: Apesar de haver bastante informação, em inglês acerca da forma como estão a ser usadas as tecnologias que apoiam a Web Semântica, em Portugal apenas consegui encontrar, informação com qualidade na Linguateca, apesar da mesma ser direccionada para os eventos HAREM.

- Importação dos dados: Os ficheiros distribuídos pelo pólo da Faculdade de Ciências de Lisboa, XLDB, que constituem a geo-ontologia Geo-Net-PT02 são ficheiros de texto com um tamanho físico entre 500 e 800 MB o que tornou incomportável a tentativa de fazer um *parser* para pesquisa dentro destes ficheiros. A solução veio com a descoberta do servidor Virtuoso que me permitiu importar os dados desses ficheiros para dentro de uma base de dados semântica. No entanto, esta importação não foi fácil devido ao volume de dados.

- Limitações do sistema: Apesar de conseguir o objectivo básico do trabalho que foi o reconhecimento de relações entre entidades geográficas, sei que o sistema ainda contém falhas, não conseguindo reconhecer algumas relações existentes nos textos. Isto deve-se em parte à grande ambiguidades que temos na língua portuguesa, sendo no entanto um dos trabalhos futuros a realizar por mim, a optimização dos reconhecimentos.

4.3.Trabalhos Futuros

- **Reconhecimento de EM identificadas mas fora de contexto**
  Para exemplificar este tipo de problema, tomemos em conta a seguinte frase "*Portugal triste com a saída do mundial* ". Nesta frase, o REMPT consegue identificar Portugal como sendo uma entidade geográfica, o que está correcto, o que não está correcto é que no contexto da frase, Portugal não se refere ao país em si, mas sim ao seu povo. O REMPT não é capaz de identificar este tipo de relação. A solução para este problema, passa pela análise sintáctica da frase, do princípio ao fim, de modo a poder reconhecer o contexto da frase.
- **Avaliação com colecção dourada:** O REMPT necessita de ser testado com a Colecção Dourada dos textos do HAREM, pois são textos anotados manualmente contendo as entidades bem identificadas e que servem para avaliarmos a percentagem de erros de identificação, classificação e de relações que existem no sistema.
  Esta avaliação irá dar-me pistas para que possa melhorar o sistema de modo a conseguir reconhecer mais entidades e com mais precisão.
- **Incorporação de Geo-ontologia Mundial:** A incorporação no sistema de uma geo-ontologia mundial irá permitir o reconhecimento de entidades em Portugal e a nível internacional. Das geo-ontologias mundiais encontradas optei por incorporar a Geonames (Geonames.org, 2010) que contém referências para entidades geográficas de todos os países.
- **Aplicações Práticas:** Faz parte também dos meu planos aplicar a geo-ontologia usada para este trabalho numa aplicação prática, relacionada com o turismo em Portugal. Para isso pretendo construir uma ontologia que liga com a existente de modo a poder cruzar dados e responder a perguntas como:
  "Qual o prato típico da localidade x e quais os restaurantes recomendados até 20 €?"

Bibliografia

Amitay, E., Har'El, N., Sivan, R.S. & Soffer, A., 2004. Web-a-Where: Geotagging Web content. In Sanderson, M., Jarvelin, K., Allan, J. & Bruza, P., eds. *Proceedings of the 27 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Sheffield, Reino Unido, 25-29 de Julho, 2004. ACM Press.

Answers.com, 2010. *Answers.com*. [Online] Disponível em: http://www.answers.com [Acedido em 05 Julho 2010].

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities. *Scientific American*, Maio. pp.284(5):34-43.

Bick, E., 2000. *The parsing systema "Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Tese de Doutoramento ed. Aarhus, Dinamarca: Aarhus University Press.

Bruckschen, M., Souza, J.G.C.d., Vieira, R. & Rigo, S., 2008. Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas. In C. Mota & D. Santos, eds. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas:O Segundo HAREM*. Linguateca. pp.247-60.

Cardoso, N., 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In C. Mota & D. Santos, eds. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca. pp.195-211.

Chaves, M.S., 2005. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In C. Mota & D. Santos, eds. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca. pp.231-45.

Chaves, M.S., 2009. *Uma Metodologia para Construção de Geo-Ontologias*. Doutoramento em informática, Especialidade de Engenharia Infomática. Lisboa: Faculdade de Ciências Faculdade de Ciências, Departamento de Informática.

Chaves, M.S., Silva, M.J. & Martins, B., 2005. A geographic knowledge base for Semantic Web applications. In C.A. Heuser, ed. *Proceedings do 20º Simpósio Brasileiro de Banco de Dados (SBBD)*. Uberlândia,MG,Brasil, 3-7 de Outubro. pp.40-54.

Cowie, J. & Wilks, Y., 2000. I*nformation Extraction. A Handbook of NaturalLanguage Processing: Techniques and Applications for the Processing of Language as Text*. Nova Iorque, EUA.

Delboni, T.M., 2005. *Expressões de posicionamento cmo fonte de contexto geográfico na web*. Dissertação de Mestrado ed. Universidade Federal de Minas Gerais - UFMG.

Desnsham, I. & Reid, J., 2003. A geo-coding service encompassing a geo-parsing tool and integrated digital gazeteer service. In Kornai, A. & Sundheim, B., eds. *Workshop on the Analysis of Geographic References at the NAACL-HLT 2003 conference*. Emonton, Canada, 2003.

Fürnkranz, J., 2002. Round Robin Classification. *Journal of Machine Learning Research*, pp.2:721-747.

Geonames.org, n.d. *Geonames*. [Online] Disponível em : http://www.geonames.org/.

Gillam, R., 1999. *Text Boundary Analysis in Java*. [Online] IBM Corp. Disponível em : http://icu-project.org/docs/papers/text_boundary_analysis_in_java/ [Acedido em 30 Junho 2010].

Google, 2008. *The official Google Blog*. [Online] Disponível em: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html [Acedido em 20 Junho 2010].

Himmelstein, M., 2005. Local Search: The Internet Is The Yellow Pages. *Computing Practices*, Fevereiro. pp.26-34.

Linguateca, 2007. *Reconhecimento de entidades mencionadas em português*. Linguateca.

Lopez-Pellicer, F.J., Chaves, M., Rodrigues, C. & Silva, M.J., 2009. Geographic Ontologies Production. In U.d. Lisboa & F.d. Ciências, eds. *Grease-II Techincal Report*. Lisboa, Portugal: LASIGE. pp.09-18.

Manguinhas, H.M.Á., Martins, B.E.d.G. & Borbinha, J., 2008. A geo-temporal Web gazetteer service integrating data from multiple sources. In *3rd IEEE International Conference on Digital Information Management*. Londres, Reino Unido: IEEE.

Marktest, 2009. *Marktest Barem Internet*. [Online] Disponível em: http://www.marktest.pt/internet/default.asp?c=1294&n=1860 [Acedido em 01 Julho 2010].

Martins, B., 2008. O sistema CaGE no Segundo HAREM. In C. Mota & D. Santos, eds. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca. pp.149-58.

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

OpenLink Software, 2010. *OpenLink Software*. [Online] Disponível em: http://virtuoso.openlinksw.com/ [Acedido em 13 Junho 2010].

Tauberer, J., 2010. *SemWeb.Net:Semantic Web /RDF Library for C#/.NET*. [Online] Disponível em : http://razor.occams.info/code/semweb/ [Acedido em 30 Maio 2010].

W3C, 2004. *Resource Description Framework (RDF):Concepts and Abstract Syntax*. [Online] Disponível em : http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ [Acedido em 16 Abril 2010].

W3C, 2008b. *SPARQL Query Language for RDF*. [Online] Disponível em: http://www.w3.org/TR/rdf-sparql-query/ [Acedido em 05 Julho 2010].

Wikipedia, 2004f. *OWL*. [Online] Disponível em : http://pt.wikipedia.org/wiki/OWL [Acedido em 10 Julho 2010].

Wikipedia, 2009. *List of Languages by number of native speakers*. [Online] Disponível em : http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers [Acedido em 20 Junho 2010].

Wikipedia, 2010a. *Information Retrieval*. [Online] Disponível em: http://en.wikipedia.org/wiki/Information_retrieval [Acedido em 14 Março 2010].

Wikipedia, 2010b. *Notation3*. [Online] Disponível em: http://en.wikipedia.org/wiki/Notation_3 [Acedido em 10 Junho 2010].

Wikipedia, 2010c. *Web semântica*. [Online] Disponível em: http://pt.wikipedia.org/wiki/Web_sem%C3%A2ntica [Acedido em 1 Junho 2010].

Wikipedia, 2010d. *Message Understanding Conference*. [Online] Disponível em: , http://en.wikipedia.org/wiki/Message_Understanding_Conference [Acedido em 1 Junho 2010].

Wikipedia, 2010e. *Ontologia (ciência da computação)*. [Online] Disponível em: http://pt.wikipedia.org/wiki/Ontologia_%28ci%C3%AAncia_da_computa%C3%A7%C3%A3o%29 [Acedido em 01 Julho 2010].

Anexo A

Tecnologias Utilizadas

Tecnologias Utilizadas

RDF

A *Resource Description Framework* (RDF), é definida, pela W3C como sendo uma framework para representação de informação na Web (W3C, 2004).

Esta Framework é um standard da W3C para a modelação e partilha de conhecimento distribuído, baseado no conceito de um mundo descentralizado. A ideia por detrás desta Framework é de que qualquer coisa, acerca de qualquer outra coisa, pode ser decomposta no que se chamam triplas.

As triplas são conjuntos de dados constituídos por um sujeito, um predicado e um objecto, que representam uma parte do conhecimento global acerca de um determinado assunto, por exemplo, na frase "Lisboa é em Portugal" podemos definir a tripla deste modo:

Sujeito: Lisboa

Predicado: é em

Objecto: Portugal

Embora tenha semelhanças com os modelos de diagrama de classes e com o modelo ER (*Entity-Relation*), o RDF é um modelo abstracto de dados com vários formatos de serialização ou seja, vários formatos de ficheiros, o que faz com que, dependendo do formato escolhido, assim sejam definidas as triplas. Esta Framework é a base, em termos de modelo de dados, do que se denomina Web Semântica.

Existem dois tipos de ficheiros mais comuns para representar o conhecimento através de triplas, o primeiro é baseado no formato XML mas sobressaindo a diferença entre este formato e o do RDF, pois este ultimo tem o MIME TYPE "application/rdf+xml" (RFC 3870). A figura 8 mostra um exemplo da composição de um elemento RDF.

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description
rdf:about="http://en.wikipedia.org/wiki/Tony_Benn
">
   <dc:title>Tony Benn</dc:title>
   <dc:publisher>Wikipedia</dc:publisher>
  </rdf:Description>
</rdf:RDF>
```

Fig. 6 –Exemplo de um elemento RDF

Neste exemplo está representado o artigo http://en.wikipedia.org/wiki/Tony_Benn, representado pelo atributo *rdf:about* , com o título Tony Benn, representado pelo elemento *dc:title* e publicado pela Wikipedia, representado pelo elemento *dc:Publisher.*

O outro tipo de ficheiro é o baseado na notação 3, ou mais simplesmente N3 (Wikipedia, 2010b), também definido pela W3C e que foi criado para permitir a anotação de triplas manualmente e para ser mais legível e compreensível para o ser humano.

O mesmo exemplo que foi dado acima para o formato RDF+XML, pode ser representado usando o formato N3 o que resulta na tripla mostrada na figura 9.

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.

<http://en.wikipedia.org/wiki/Tony_Benn>
  dc:title "Tony Benn";
  dc:publisher "Wikipedia".
```

Fig. 7 – Exemplo de um elemento N3

OWL

Com o crescente uso de ontologias, nomeadamente no que à Web Semântica diz respeito, surgiu também a necessidade de as conseguir ligar entre elas, de modo a poder processar informação referente a vários domínios. Deste modo surge a *Web Ontology Language* (OWL), uma linguagem para definir e instanciar ontologias na Web, tendo sido aceite como padrão, pela W3C em 2004 (Wikipedia, 2004f).

Existem 3 sub-linguagens associadas a esta linguagem:

- ***OWL Lite*** : Suporta apenas relações de 0 ou 1 em termos de cardinalidade e tem uma classificação hierárquica, sendo a mais simples das formas de implementação;
- ***OWL DL :*** È a forma mais usada desta linguagem, pois garante o uso de todas as classes da linguagem, mantendo ao mesmo tempo a capacidade de computação obrigando a que todas as operações terminem em tempo finito;
- ***OWL Full :*** *Dá* acesso a todos os recursos da linguagem mas, por outro lado terá de ser sacrificada a parte computacional, pois esta sub-linguagem faz uso total do RDF sem garantia de acabar o processamento, muito menos em tempo útil.

SPARQL

Com o surgimento das ontologias, foi definido o padrão que permite pesquisa dentro dessas ontologias, o SPARQL (W3C, 2008b), um acrónimo para *SPARQL Protocol and RDF Query Language.* A sintaxe usada pelo SPARQL para consulta de base de dados de ontologias é semelhante à usada com o SQL, sendo fácil, para quem já tenha trabalhado com SQL, começar a fazer pesquisas com SPARQL. Por exemplo, se quisermos procurar na ontologia usada neste trabalho, pelos dados relativos a Lisboa podemos usar a seguinte pesquisa:

```
select * where {?entity dcterms:title "Lisboa"@pt . ?entity ?type ?value}
```

O resultado obtido é demasiado extenso, pelo que mostro aqui apenas uma pequena parte desse resultado.

entity	type	value
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net#Feature
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://www.w3.org/2000/01/rdf-schema#label	Lisboa (Concelho)
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://purl.org/dc/terms/title	Lisboa
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://purl.org/dc/terms/alternative	1106
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net#inDomain	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#GeoAdministrative
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net#type	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#concelho-ATCON
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net#lineage	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#SRC-INE
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt#preferred	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-pt
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt#isPartOf	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF3965
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt#isPartOf	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#estremadura-AF418732
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt#isPartOf	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#grande_lisboa-AF129
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net#footprint	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#AFP4471
http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#lisboa-AF146	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt#identifier	http://xldb.di.fc.ul.pt/xldb/publications/2009/10/geo-net-pt-02#1106-DICOFRE

Tabela 5 – Resultado da pesquisa por "Lisboa" usando SPARQL

Este é o aspecto normal de uma consulta à base de dados de uma ontologia. Neste caso escolhi como saída de resultados o formato HTML por ser mais fácil de

entender. A partir do resultado obtido podemos mostrar informação sobre Lisboa, como, as ruas, as freguesias, as localidades adjacentes, etc.

Virtuoso

O tratamento de grandes volumes de informação não pode, actualmente, ser feito em tempo útil, sem a ajuda de um motor de base de dados. No caso das ontologias, a informação pode crescer a um ritmo acelerado, dependendo do tema em questão, o que é o caso da geo-ontologia.

O Virtuoso (OpenLink Software, 2010) é definido como sendo um servidor de base de dados SQL, objecto-relacional, de alto desempenho. Este servidor permite-nos a criação de bases de dados semânticas, baseadas em RDF ou em N3 de modo a podermos fazer pesquisas usando SPARQL. Para podermos usar o SPARQL, o servidor cria automaticamente um SPARQL *endpoint*, apontado para o URI que escolhermos quando criamos a nossa base de dados. Para além do *endpoint* é também disponibilizado de forma automática um WebService. O Virtuoso permite-nos ainda escolher o formato de saída dos dados entre HTML, XML, JSON ou mesmo para uma folha EXCEL.

Distribuído pelo consórcio OpenLink Software como software Open-Source, vem facilitar o uso das tecnologias associadas à Web Semântica, elas próprias distribuídas como formatos Open-Source.

SemWeb.NET

A biblioteca SemWeb (Tauberer, 2010), é uma biblioteca que permite a interpretação das tecnologias usadas para a Web Semântica, como o RDF ou N3 através do uso de SPARQL. Esta biblioteca foi criada para a linguagem de programação C# que faz parte da Framework NET da Microsoft.