Capítulo 12

REMMA - Reconhecimento de Entidades Mencionadas do MedAlert

Liliana Ferreira, António Teixeira e João Paulo da Silva Cunha

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 12, p. 213–229.

Este capítulo descreve o sistema REMMA (Reconhecimento de Entidades Mencionadas do MedAlert), um reconhecedor de entidades mencionadas que usa a Wikipédia como fonte de conhecimento externo. O REMMA foi desenvolvido no âmbito do projecto MedAlert – Sistema de Processamento de Linguagem Médica (http://www.ieeta.pt/sias/medalert). O MedAlert usa a informação disponibilizada pela Rede Telemática de Saúde (RTS) (Cunha et al., 2006), em utilização no Hospital Infante D. Pedro e na região de Aveiro, e tem como principal objectivo a utilização de técnicas de extracção automática de informação de textos médicos, de modo a inferir, de uma forma automática, irregularidades/dúvidas suscitadas pelas decisões tomadas pelos profissionais de saúde. O MedAlert, que deverá tomar a forma dum módulo escalável e adaptável a diferentes configurações de sistemas de informação hospitalares, pretende usar técnicas de processamento de linguagem natural (PLN) para extrair informação de um amplo conjunto de textos médicos, particularmente cartas de alta e textos contendo directivas médicas. Esta informação, bem como a proveniente de recursos externos como ontologias e outras fontes de conhecimento médico, deverá ser utilizada no apoio e validação de decisões, melhorando, assim, o cuidado médico, com a redução de erros, melhoria de segurança e aumento da satisfação. O REM é considerado como uma subtarefa importante da maioria das aplicações de engenharia de linguagem e um primeiro passo para a extracção de informação. Deste modo, tornou-se essencial o desenvolvimento de um módulo capaz de identificar e classificar entidades que respondam a um conjunto de perguntas relevantes, reduzindo, consequentemente, a complexidade da extracção de factos.

O REMMA é apresentado neste capítulo no âmbito da sua participação no Segundo HAREM e, consequentemente, como um sistema de REM em textos não especializados. Esta participação teve como objectivo principal explorar diferentes abordagens e perceber qual a utilidade da utilização de fontes de conhecimento externo na tarefa de REM, para uma posterior adaptação à área em que nos concentramos, a medicina.

O REM foi definido nas conferências MUC (Hirschman, 1998) como sendo a tarefa de detectar e classificar expressões em texto que pertençam a diferentes classes (por exemplo, pessoa, local, organização, data, tempo). Desde que o REM apareceu, duas principais aproximações foram adoptadas para lidar com a tarefa. Uma é referida como baseada em conhecimento e usa explicitamente recursos tais como regras e almanaques construídos e mantidos, de uma forma geral, manualmente. A outra segue o paradigma da aprendizagem automática e usa normalmente como colecção de treino um corpo anotado que é usado para o treino de um algoritmo de aprendizagem supervisionada. Inicialmente, e principalmente para as conferências MUC, a maior parte dos sistemas REM usavam uma aproximação baseada em conhecimento. Este método provou obter bons resultados, tendo, o melhor sistema obtido, na medida F, uma classificação de 0,9339 (Mikheev et al., 1998). No entanto, esta aproximação apresenta um problema relevante: os almanaques e as regras são difíceis de construir e manter, sendo, em particular, difícil evitar a sobreposição entre almanaques.

O REMMA tenta contornar esta questão através da utilização da Wikipédia como fonte de conhecimento externo, em particular, através da extracção de categorias semânticas a partir da primeira frase de uma página da Wikipédia.

12.1 A Wikipédia como fonte de conhecimento para REM

Recentemente, tem-se vindo a assistir a um crescimento rápido e bem-sucedido da Wikipédia (http://www.wikipedia.org), uma enciclopédia electrónica livre e que está a ser construída por milhares de colaboradores em todo o mundo. A Wikipédia tinha em Outubro de 2008 mais de 2 561 000 artigos na versão inglesa e cerca de 428 000 artigos na sua versão portuguesa. Uma vez que a Wikipédia pretende ser uma enciclopédia, a maior parte dos artigos são sobre entidades mencionadas e mais estruturados do que texto livre. A Wikipédia é actualizada diariamente, ou seja, novas entidades são adicionadas e revistas constantemente (Voss, 2005). Deste modo, a extracção de conhecimento a partir da Wikipédia para o PLN é uma forma promissora de permitir a criação de aplicações em grande escala, aplicáveis em situações da vida real. De facto, vários estudos surgiram recentemente em que a Wikipédia é explorada como fonte de conhecimento (Auer et al., 2007; Ruiz-Casado et al., 2006; Santos et al., 2008a; Wu e Weld, 2007; Zesch et al., 2008). A maior parte destes estudos concentram-se na extracção automática de almanaques da Wikipédia (Toral e Muñoz, 2006) e na utilização da estrutura interna da Wikipédia para a desambigação de entidades mencionadas (Bunescu e Pasca, 2006). O estudo com mais relevância para o trabalho apresentado neste capítulo é o de Kazama e Torisawa (2007), onde se utiliza o sintagma nominal da primeira frase de um artigo da Wikipédia para a extracção da categoria semântica. No REMMA, optou-se por identificar na primeira frase do artigo um conjunto de palavras indicativas da categoria e tipo de uma dada entidade. Com este trabalho pretende-se determinar até que ponto as classificações semânticas extraídas a partir de um artigo da Wikipédia, em particular da primeira frase do artigo, podem ser consideradas como *definições* da entidade descrita no artigo. Por exemplo, o artigo da Wikipédia sobre a Universidade de Aveiro começa com a frase (12.1).

(12.1) *A Universidade de Aveiro (UA) é uma universidade pública portuguesa localizada em Aveiro.*

A extracção da palavra *universidade* desta frase permite inferir a classificação a atribuir à entidade *Universidade de Aveiro*. O método utilizado na obtenção destas classificações é descrito em detalhe na secção 12.2.

Várias razões determinaram a escolha da Wikipédia para utilização como fonte de informação no REMMA. A principal foi a necessidade de desenvolver um sistema capaz de reconhecer entidades de domínio geral e a impossibilidade de construir ou aceder a um almanaque de grande dimensão. Outras motivações baseiam-se em diversas características da Wikipédia, como por exemplo:

- É um recurso de informação de grandes dimensões. Em Outubro de 2008 continha mais de 7 milhões de artigos em cerca de 200 línguas e aproximadamente 428 mil entradas na versão portuguesa.
- O seu conteúdo tem uma licença livre, estando sempre disponível para a investigação sem restrições e sem a necessidade da aquisição de direitos.
- É um recurso de domínio geral, podendo, desta forma, ser usado na tarefa de extracção de informação de sistemas de domínio aberto.

- Os dados apresentados têm algum grau de formalidade e de estruturação (por exemplo, categorias) o que ajuda no seu processamento.
- É actualizada e revista continuamente através da colaboração de diversas pessoas.

A Wikipédia disponibiliza todo o conteúdo para cada uma das diferentes línguas, em formato XML, bem como as ferramentas necessárias para a sua conversão para SQL. O REMMA utiliza a informação disponibilizada em formato SQL e fez uso da estrutura interna desta base de dados. As secções seguintes descrevem a estrutura básica da Wikipédia no seu contexto da sua utilização no REMMA. O esquema completo da base de dados pode ser consultado em `http://www.mediawiki.org/wiki/Manual:Database_layout`.

12.1.1 Estrutura básica

Uma página da Wikipédia é identificada por um nome único, que pode ser obtido através da concatenação das palavras existentes no título com "_", mantendo a primeira letra da primeira palavra maiúscula. Seguindo o exemplo anterior, o nome único para a página *Universidade de Aveiro* é `Universidade_de_Aveiro`.

Usualmente, o título da página é o nome mais comum para a entidade descrita neste. Quando o nome é ambíguo, o título é também qualificado com uma expressão em parênteses, como no caso da página referente à flor *Cravo*, que é descrita, na Wikipédia, na página intitulada `Cravo_(flor)`.

De uma forma geral, existe um relacionamento de correspondência de muitos-para-muitos entre os nomes e as entidades. Este relacionamento é definido na Wikipédia através das *páginas de redirecção* e das *páginas de desambiguação*. Estes dois conceitos são explorados em mais detalhe nas secções 12.1.2 e 12.1.3.

No entanto, as páginas da Wikipédia têm outras estruturas úteis para a extracção de conhecimento, tais como as categorias e as ligações internas. Estes dois conceitos são descritos nas secções 12.1.4 e 12.1.5.

12.1.2 Redirecção

Existe uma *página de redirecção* para cada nome alternativo que possa ser usado para referir uma entidade na Wikipédia. As redirecções são marcadas como `#REDIRECT [[A B C]]` nos ficheiros fonte, onde "`[[...]]`" é a sintaxe que indica uma ligação a outro artigo na Wikipédia. As redirecções são usadas por várias razões relacionadas com a ambiguidade. Por exemplo, são usadas para expansão de abreviaturas tal como de `UA` para `Universidade de Aveiro`. Também são usadas no contexto de desambiguações mais difíceis, como as descritas na secção seguinte.

12.1.3 Páginas de desambiguação

Alguns autores criam uma *página de desambiguação* para um nome de entidade ambíguo[1]. Estas páginas são usadas para nomes que podem ter vários significados e possuem referências a outras páginas que dizem respeito a diferentes entidades que partilham o mesmo nome, enumerando todos os artigos possíveis para esse nome. Por exemplo, a página de

[1] *Ambíguo* refere-se ao caso em que o nome pode ser usado para referir várias entidades (i.e., artigos da Wikipédia)

desambiguação para o nome *Madeira* lista doze entidades associadas, isto é, para além dos nomes não ambíguos originados pelas páginas de redirecção, pode encontrar-se nestas páginas outros homónimos de uma dada entidade. A tabela 12.1 apresenta alguns exemplos das páginas que são apresentadas através da desambiguação do nome *Madeira*.

12.1.4 Categorias

Toda a página da Wikipédia deve ter pelo menos uma categoria. A categoria é uma página especial gerada automaticamente a partir das ligações que a esta vão dar. Regra geral, e para fins de organização, toda e qualquer página da Wikipédia deve ser categorizada por quem a criou de modo a garantir a geração automática da página da categoria e uma correcta catalogação das páginas da Wikipédia.

A tabela 12.1 apresenta alguns exemplos que exploram a organização interna da Wikipédia. Por exemplo, a página sobre o arquipélago da Madeira, com o título *Madeira (arquipélago)*, está associado a um conjunto de categorias, entre as quais *Região Autónoma da Madeira* e *Regiões vitivinícolas*.

Tabela 12.1: Exemplos de títulos e categorias de artigos relativos à desambiguação da palavra *Madeira*.

Título	**Redirecção**	**Categorias**
Madeira (material)	Madeira	Madeira
Madeira (arquipélago)	Região Autónoma da Madeira	Região Autónoma da Madeira, NUTS III portuguesas, ... Regiões vitivinícolas
Madeira Beach	—	Cidades da Flórida
Jamila Madeira	—	Loulé, Políticos Portugal
Vinho Madeira	—	Vinhos de Portugal

12.1.5 Ligações internas

Os artigos da Wikipédia contêm frequentemente menções a entidades já definidas. Estas ligações devem ser feitas através da utilização de ligações internas. Dois exemplos de ligações internas estão representados no exemplo (12.2) retirado da página sobre a Região Autónoma da Madeira.

(12.2)
```
A Madeira, oficialmente designada por Região Autónoma da
Madeira, é um território [[Portugal|português]] dotado de
autonomia política e administrativa através do [[Estatuto
Político Administrativo da Região Autónoma da Madeira]],
previsto na [[Constituição da República Portuguesa]].
```

A expressão da segunda ligação (*Estatuto Político Administrativo da Região Autónoma da Madeira*) corresponde ao título do artigo a que se refere. A mesma expressão é usada na

versão apresentada ao utilizador. Se o autor quiser que seja apresentada uma expressão diferente (por exemplo, *português* em vez de *Portugal*) então a expressão alternativa é incluída numa ligação com outro nome (em inglês, *piped link*), após o titulo. O exemplo (12.3) ilustra a expressão apresentada para o exemplo anterior.

(12.3) `A Madeira, oficialmente designada por Região Autónoma da Madeira, é um território português dotado de autonomia política e administrativa através do Estatuto Político Administrativo da Região Autónoma da Madeira, previsto na Constituição da República Portuguesa.`

12.2 O sistema REMMA

Nesta secção é descrito em mais detalhe o sistema REMMA e a sua arquitectura. A secção 12.2.1 apresenta a plataforma base usada pelo REMMA. A secção 12.2.2 foca a arquitectura do sistema, descrevendo os métodos utilizados para a classificação das entidades.

12.2.1 A plataforma base - UIMA

Uma característica do sistema é a sua integração na plataforma UIMA. O UIMA, *Unstructured Information Management Architecture* (Ferrucci e Lally, 2004), é uma plataforma livre, escalável e extensível, para a criação, integração e desenvolvimento de sistemas de gestão de informação não estruturada. Embora seja uma arquitectura com um certo grau de complexidade, tem diversas vantagens, como por exemplo:

- Disponibiliza algumas ferramentas de pré-processamento, tais como leitores e finalizadores genéricos, atomizador, separador em frases e outros anotadores simples;
- Uniformiza a estrutura dos resultados;
- Foca a modelação em vez de na programação.

O UIMA usa uma Estrutura de Análise Comum (em inglês, *Common Analysis Structure*, CAS) que permite aos anotadores acesso de leitura ao objecto a ser processado (por exemplo, um documento) e acesso de leitura/escrita aos resultados da análise ou às anotações associadas às diferentes regiões dos objectos. Estas regiões podem corresponder a palavras, frases ou parágrafos no texto. A CAS é partilhada entre os diversos anotadores que processam a colecção de objectos, passando de um anotador para o seguinte no processo.

12.2.2 A arquitectura

A arquitectura do REMMA está apresentada na figura 12.1.

O REMMA começa por ler os documentos, um por um, e guardar os respectivos metadados. No caso da colecção do Segundo HAREM é guardada a identificação do documento em análise. Os textos são posteriormente divididos em frases e átomos com a ajuda das ferramentas de pré-processamento disponíveis no UIMA. O analisador TreeTagger (Schmid, 1995) foi usado na obtenção das categorias morfossintácticas. Este analisador morfossintáctico foi utilizado exclusivamente para eliminar algumas preposições e advérbios dos candidatos a EM, justificando-se desta forma o uso de um analisador estatístico.

Figura 12.1: Arquitectura do REMMA

As anotações geradas por estas ferramentas são armazenadas na CAS e usadas nos diversos anotadores que constituem o módulo de REM. A figura 12.2 apresenta a sequência de anotadores utilizados na identificação e classificação das entidades.

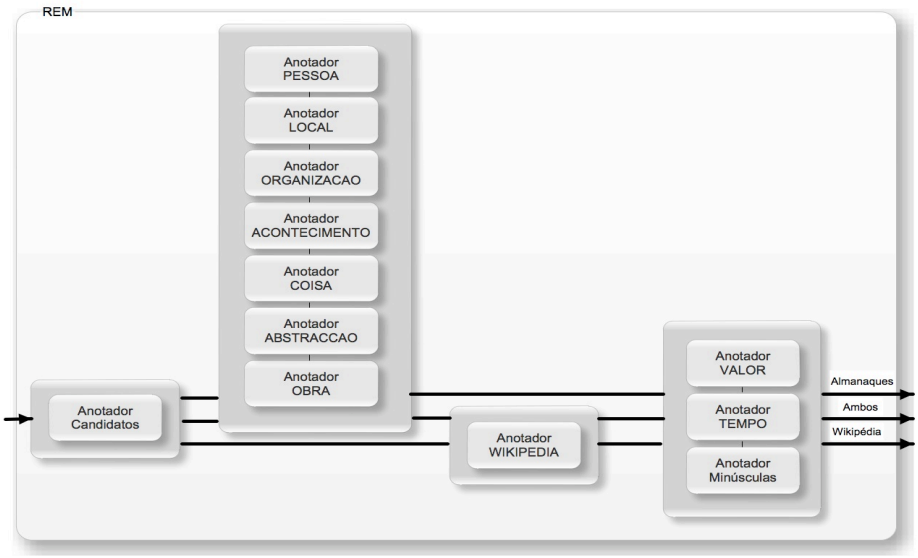

Figura 12.2: Anotadores do REMMA

O primeiro anotador a ser invocado é o `Anotador de Candidatos` que identifica todas as expressões candidatas a entidades mencionadas. As entidades candidatas são todos os conjuntos de termos iniciados por letra maiúscula. Na geração da expressão candidata foi também considerada a presença de termos de ligação com um comprimento inferior a 5 caracteres (por exemplo, *e*, *em*, *de*, *da*, *do*, *dos*, *das*, *para*, etc.). Não foram contempladas expressões contendo algarismos. Estas expressões candidatas foram posteriormente analisadas pelos anotadores de classificação.

O REMMA foi desenvolvido de modo a contemplar duas abordagens de classificação distintas. A primeira baseia-se em almanaques e regras muito simples e é descrita na secção seguinte. A classificação com base na informação extraída da Wikipédia pode ser realizada em conjunto ou separadamente do método anterior e é apresentada em detalhe na secção 12.2.2.2 . A utilização de duas abordagens distintas justifica-se pela necessidade de perceber quais as vantagens e desvantagens inerentes a cada método, em particular, de que forma a utilização da Wikipédia permite melhorar os resultados.

Na tarefa de classificação com base na informação extraída da Wikipédia foi utilizado um subconjunto de todo o conteúdo da Wikipédia, que é disponibilizado em XML para cada uma das diferentes línguas. Foi utilizada a Wikipédia portuguesa de Fevereiro de 2008, que inclui 1 290 836 páginas. Os dados foram posteriormente exportados para uma base de dados SQL, de modo a poderem ser usados neste sistema. Optou-se por não usar a informação existente nas páginas de desambiguação, mas apenas a redirecção que a comunidade Wikipédia entende ser a que mais utilizadores estão à procura. Relativamente às categorias associadas a cada página da Wikipédia, observou-se que uma página pode ter mais do que uma categoria, e que muitas vezes estas categorias não são claros hiperónimos da entidade a ser analisada. Assim, esta informação não foi usada, uma vez que a sua utilização implicaria a necessidade de seleccionar uma categoria apropriada nas categorias listadas, ficando esta tarefa para trabalho futuro.

Os anotadores relativos às categorias `TEMPO` e `VALOR` e o anotador `Minúsculas`, desenvolvido para a inclusão nas entidades da informação relativa às palavras começadas por minúsculas contempladas nas directivas do Segundo HAREM, são descritos separadamente na secção 12.2.2.3.

12.2.2.1 Classificação com base em regras e almanaques

Esta primeira abordagem baseou-se numa utilização combinada de um conjunto de regras de análise de contexto com a consulta de diversos almanaques de pequena dimensão. Os almanaques utilizados tinham sido já criados manualmente no âmbito de projectos desenvolvidos anteriormente na área da extracção de informação de relatórios médicos e contêm nomes de entidades de diversas classes semânticas, como por exemplo, listas de nomes de pessoas, listas de cidades portuguesas, listas de doenças e sintomas clínicos, etc.

As regras utilizadas foram criadas manualmente e baseiam-se no contexto em que a expressão é referida. Estas regras exploram certas classes semânticas de palavras, como por exemplo as relativas a cargos, tipos de locais, tipos de organização e outros. A tabela 12.2 lista alguns exemplos de palavras utilizadas para anotar as classes semânticas `PESSOA`, `LOCAL`, `ORGANIZACAO` e `ACONTECIMENTO`. Note-se que a tabela apenas apresenta informação relativa à anotação da *categoria* da entidade. Estas listas foram posteriormente subdivididas de modo a fornecerem informação relativa à anotação *tipo* da entidade mencionada, caso esta exista.

Os anotadores que usam a informação contida nestes almanaques e regras começam por dividir a expressão candidata nos seus vários termos e atribuem uma categoria semântica caso algum dos termos da expressão exista nas listas usadas. Quando esta anotação não é conseguida, procuram na expressão candidata palavras pertencentes à classe semântica em análise. Dois exemplos ilustrativos do tipo de cobertura deste módulo são apresentados em (12.4), onde *Paulo* é um nome existente no almanaque relativo a nomes

Tabela 12.2: Exemplos e quantidade de palavras usadas para a definição de regras contextuais das entidades PESSOA, LOCAL, ORGANIZACAO e ACONTECIMENTO.

PESSOA (N=110)	**LOCAL** (N=58)	**ORGANIZACAO** (N= 50)	**ACONTECIMENTO** (N=28)
Ministro	Praça	Museu	Campeonato
Chefe	Avenida	Faculdade	Concerto
Princesa	Rua	Sindicato	Congresso
Reitora	Cidade	Prefeitura	Exposição
...	...	...	...

de pessoas, e (12.5), onde *Provedor* é uma das palavras usadas nas regras contextuais da entidade PESSOA.

(12.4) **Entrada:** Ao que parece, Paulo Pinto Mascarenhas tem a convicção firme
Saída: Ao que parece, <EM ID="xxx" CATEG="PESSOA" TIPO="INDIVIDUAL">**Paulo Pinto Mascarenhas**</EM> tem a convicção firme

(12.5) **Entrada:** do Provedor do Espectador, não veio qualquer espécie de pressão.
Saída: do <EM ID="xxx" CATEG="PESSOA" TIPO="CARGO">**Provedor do Espectador**</EM>, não veio qualquer espécie de pressão.

12.2.2.2 Classificação com recurso à Wikipédia

A classificação com base na informação extraída da Wikipédia pode ser realizada em conjunto com a descrita na secção anterior, analisando neste caso apenas as entidades candidatas não anotadas anteriormente, ou individualmente, procurando uma classificação para todas as entidades candidatas identificadas.

Aquilo que se pretende é que este anotador seja capaz de encontrar uma entidade na Wikipédia correspondente à identificada nos textos em análise. Por exemplo, na frase (12.6), a expressão candidata a EM *Universidade de Harvard* é identificada pelo Anotador de Candidatos (ver secção 12.2.2). O objectivo passa por perceber de que forma é que esta entidade é descrita na Wikipédia e, consequentemente extrair a respectiva classificação do artigo.

(12.6) Concluiu os seus estudos de medicina, em 1870, na Universidade de Harvard, onde iniciou a sua carreira como professor de fisiologia em 1872.

Deste modo, cada uma das entidades candidatas identificadas é convertida num identificador da Wikipédia através da concatenação dos vários termos da expressão com o caracter "'_"'. Por exemplo, a expressão *Universidade de Harvard* é convertida em Universidade_de_Harvard e o artigo correspondente recuperado, seguindo a redirecção, caso esta exista, até obter uma página de não-redireccionamento[2].

[2] Existem na Wikipédia algumas páginas para outros conteúdos que não os usuais artigos. Estes são distinguidos por um atributo *namespace*. Para a recuperação dos artigos que precisamos foi apenas analisado o *namespace* 0, que é o mais comum para estes artigos.

Embora não exista uma regra de formatação estrita, é normal que os artigos da Wikipédia comecem com uma pequena frase que define a entidade descrita no artigo. Por exemplo, o artigo com o título `Universidade_de_Harvard` começa com a frase (12.7).

(12.7) A Universidade Harvard (em inglês Harvard University) é uma das instituições educacionais mais prestigiadas do mundo, bem como a instituição de ensino superior mais antiga dos Estados Unidos da América.

Tal como neste exemplo, a primeira frase da maioria dos artigos contém uma expressão que indica a categoria semântica da entidade em análise. Neste caso, é a palavra *instituição*.

O método seguido concentra-se assim na extracção de tais nomes, a partir da primeira frase do artigo. Para tal foi necessário começar por remover etiquetas desnecessárias, tais como itálicos, negritos e ligações internas. As ligações internas foram convertidas para a expressão adequada (por exemplo, `[[língua inglesa | inglês]]` para `inglês`, ver secção 12.1.5). O artigo foi posteriormente dividido em frases de acordo com os padrões `\n`, `<br>` e regras simples de segmentação para o ponto final (.).

Após obtenção da primeira frase foram aplicadas regras simples, semelhantes às utilizadas no método anterior, ou seja, procuram na primeira frase do artigo da Wikipédia palavras-chave indicativas da classe semântica do artigo. Alguns exemplos, bem como a quantidade de palavras utilizadas por este anotador, são listados na tabela 12.3.

Tabela 12.3: Exemplos e quantidade de palavras-chave usadas na extracção de uma categoria semântica da primeira frase de um artigo.

PESSOA (N=15)	**LOCAL** (N=15)	**ORGANIZACAO** (N=12)	**ACONTECIMENTO** (N=2)
imperador	planeta	partido	acordo
engenheiro	cidade	movimento	competição
professor	ilha	universidade	
piloto	continente	...	
...	...	...	

Um exemplo de aplicação deste anotador, retirado da colecção usada no Segundo HAREM, é ilustrado em (12.8).

(12.8) **Entrada:** A popularidade do piloto Ayrton Senna na França era comparável à de seu maior rival, Alain Prost, quatro vezes campeão mundial.
Saída: A popularidade do piloto <EM ID="xxx" CATEG="PESSOA" TIPO="INDIVIDUAL">**Ayrton Senna**</EM> na França era comparável à de seu maior rival, <EM ID="yyy" CATEG="PESSOA" TIPO="INDIVIDUAL">**Alain Prost**</EM>, quatro vezes campeão mundial.

De notar que os nomes das entidades anotadas na frase, *Ayrton Senna* e *Alain Prost*, não existiam em nenhum dos almanaques utilizados no método anterior e também que, no caso da entidade *Alain Prost*, não existia qualquer regra contextual que a reconhecesse.

12.2.2.3 Anotadores VALOR, TEMPO e Minúsculas

Os anotadores desenvolvidos para as categorias semânticas TEMPO, TEMPO e Minúsculas são apresentados separadamente pois são independentes dos descritos anteriormente e utilizados na produção de todas as corridas do REMMA.

Os anotadores VALOR e TEMPO começam por identificar conjuntos de termos contendo pelo menos um algarismo ou que pertençam a uma lista de palavras pré-definida. No caso do anotador tempo as listas contêm, por exemplo, nomes de épocas festivas e estações do ano (*Páscoa*, *Carnaval*, *Primavera*, *Verão*, ...), dias da semana, meses, advérbios de frequência (*diariamente*, *todos os anos*, ...), etc. Para o anotador valor foram utilizadas listas contendo nomes de várias unidades (*metro*, *Kg*, *Gb*, etc.) e de nomes de moedas (*euros*, *dólares*, *contos*, etc.).

Estes anotadores incluem expressões regulares para identificar expressões como *em 25 de Abril [de 1974]*.

O anotador de minúsculas expande a anotação efectuada a uma dada entidade caso esta seja precedida por uma palavra começada por minúscula, que esteja incluída nas respectivas directivas. Caso a entidade não tenha ainda sido anotada, a palavra em minúscula precedente é analisada, de modo a inferir qual a anotação que deverá ser adicionada.

Um exemplo de aplicação, para cada um dos anotadores TEMPO e Minúsculas, encontra-se ilustrado em (12.9) e (12.10), respectivamente. Em (12.10), *(ex-)presidente* pertence à lista de palavras em minúsculas definida pelas directivas (a lista completa encontra-se no apêndice A, secção A.6).

(12.9) **Entrada:** As fortes chuvas que atingiram ontem
Saída: As fortes chuvas que atingiram <EM ID="xxx" CATEG="TEMPO" TIPO="TEMPO_CALEND">**ontem**</EM>

(12.10) **Entrada:**quando o ex-presidente José Sarney disse que sua maior missão era conduzir o país até as eleições.
Saída:quando o <EM ID="xxx" CATEG="PESSOA" TIPO="CARGO">**ex-presidente José Sarney**</EM> disse que sua maior missão era conduzir o país até as eleições.

Após a anotação das entidades identificadas pelos vários métodos descritos, um último anotador é chamado, o Finalizador. Este anotador analisa a CAS e cria o(s) documento(s) de saída. É este anotador que produz o documento XML final, através da análise das anotações associadas às diferentes regiões do(s) documento(s). É também neste passo que é efectuado o processamento de anotações alternativas <ALT>. O Finalizador determina a existência de duas ou mais anotações referentes a regiões encerradas noutra(s). Neste caso, as entidades são etiquetadas com duas ou mais anotações distintas separadas por "|". Um exemplo da saída gerada por este anotador é apresentada em (12.11).

(12.11) <ALT>
<EM ID="xxx" CATEG="PESSOA" TIPO="CARGO">**líder do Sinn Fein**</EM>
| líder do <EM ID="yyy" CATEG="ORGANIZACAO" TIPO="INSTITUICAO">**Sinn Fein**</EM>
</ALT>

12.3 Resultados no Segundo HAREM

Tal como referido, a participação do REMMA no Segundo HAREM pretendia avaliar de que forma a extracção de conhecimento a partir da Wikipédia para o REM permite criar aplicações úteis e substituir a utilização, e consequentemente a criação e manutenção, de listas e almanaques de grande dimensão. Deste modo, e tendo em consideração a arquitectura do sistema REMMA, foram geradas três corridas distintas. O nome de cada corrida é relativo à fonte de conhecimento principal usada para a obter (dentro de parêntesis encontra-se o nome que lhes foi atribuído pela organização).

- **Corrida Almanaques (REMMA_2_corr):** Corrida criada com a utilização dos anotadores de regras contextuais e almanaques. Em particular foram utilizados anotadores para as classes semânticas `LOCAL`, `PESSOA`, `ORGANIZACAO`, `ACONTECIMENTO`, `OBRA`, `ABSTRACCAO` e `COISA`. A estes anotadores seguiram-se os anotadores `VALOR`, `TEMPO` e `Minúsculas`.
- **Corrida Wiki (REMMA_3_corr):** Corrida gerada pela utilização isolada do anotador `Wikipédia`, seguido dos anotadores `VALOR`, `TEMPO` e `Minúsculas`.
- **Corrida Ambos (REMMA_1_corr):** Corrida gerada pela utilização sequencial de todos os anotadores desenvolvidos.

Relembramos da secção 12.2.2 que a figura 12.2 apresenta a sequência de anotadores utilizados na produção das diversas corridas.

As secções seguintes apresentam os resultados obtidos no Segundo HAREM na tarefa de classificação.

12.3.1 Usar a Wikipédia tem potencial para melhor desempenho?

Comparando os resultados obtidos pelas diferentes corridas, apresentados na tabela 12.4, e relativamente à medida F, podemos observar que a melhor corrida é a *Ambos*. Esta observação é verdadeira em ambas as avaliações de ALT, estrita e relaxada. No entanto, é de notar, que este resultado é obtido à custa de uma maior abrangência em relação às restantes corridas (mais $\sim$0,13 em relação à corrida *Wiki* e mais $\sim$0,07 em relação à corrida *Almanaques*, em ambas as avaliações (estrita e relaxada)) e de uma ligeira perda de precisão em relação à corrida *Almanaques* (menos $\sim$0,01).

A corrida *Wiki* é a que obtém piores resultados em todas as métricas e em ambas as avaliações. Observa-se uma diminuição dos valores da medida F em cerca de 0,12 (avaliações estrita e relaxada) em relação à corrida *Ambos*.

A utilização de almanaques gera resultados superiores aos obtidos com a utilização isolada da Wikipédia, sendo, no entanto, ainda inferiores aos obtidos com a utilização de todos os anotadores. Relativamente à corrida *Ambos*, observa-se ainda um decréscimo de aproximadamente 0,05 na medida F (ambas as avaliações).

12.3.2 Para a Wikipédia todas as categorias nascem iguais?

A tabela 12.5 apresenta os resultados obtidos para cada uma das categorias, na tarefa de classificação. Sobressai da análise da tabela o facto de as categorias `LOCAL`, `PESSOA`,

Tabela 12.4: Resultados do REMMA no HAREM clássico para a tarefa de classificação.

Versão	Avaliação estrita de ALT			Avaliação relaxada de ALT		
	Precisão	**Abrangência**	**Medida F**	**Precisão**	**Abrangência**	**Medida F**
Almanaques	**0,6132**	0,2952	0,3985	**0,6340**	0,3084	0,4150
Wiki	0,5808	0,2316	0,3312	0,5950	0,2409	0,3429
Ambos	0,6050	**0,3615**	**0,4526**	0,6226	**0,3750**	**0,4681**

ORGANIZACAO, ACONTECIMENTO e OBRA obterem claramente melhores resultados pela utilização da Wikipédia. As categorias ABSTRACCAO e COISA parecem ser imunes à utilização da Wikipédia.

Observa-se também que os resultados obtidos para as categorias VALOR e TEMPO são independentes da utilização da Wikipédia quando usada em conjunto com almanaques, tendo, no entanto, curiosamente, sofrido um ligeiro decréscimo na medida F (menos de 0,01) com a utilização isolada da Wikipédia.

Tabela 12.5: Resultados do REMMA para cada uma das categorias na tarefa de classificação.

Categoria	Melhor Versão	Classificação		
		Precisão	**Abrangência**	**Medida F**
LOCAL	Ambos	0,5700	0,5089	**0,5377**
PESSOA	Ambos	**0,6666**	0,3677	0,4740
VALOR	Ambos = Almanaques	0,3589	**0,5202**	0,4247
ORGANIZACAO	Ambos	0,5829	0,2397	0,3397
TEMPO	Ambos = Almanaques	0,4744	0,2538	0,3307
ACONTECIMENTO	Ambos	0,4044	0,1473	0,3159
OBRA	Ambos	0,5146	0,1212	0,1962
ABSTRACCAO	Ambos = Almanaques	0,2231	0,0392	0,0667
COISA	Ambos = Almanaques	0,2227	0,0318	0,0557

Na figura 12.3 observa-se a distribuição dos valores da precisão e abrangência discriminados por categoria e corrida, onde se nota uma maior precisão do REMMA relativamente à abrangência. Mais uma vez se observam os piores resultados obtidos nas categorias semânticas ABSTRACCAO e COISA.

12.3.3 Esta abordagem é competitiva?

Nesta secção tenta-se perceber até que ponto o sistema REMMA é competitivo, analisando os resultados de uma forma comparativa com os obtidos pelos melhores e piores sistemas para cada uma das categorias.

A figura 12.4 apresenta os valores obtidos em cada uma das métricas, discriminados em termos de categoria, e após uma normalização relativa aos valores máximos e mínimos obtidos, isto é, relativa aos resultados do melhor e do pior sistema em cada uma das categorias. Por exemplo, é possível perceber que o REMMA obteve a melhor precisão para

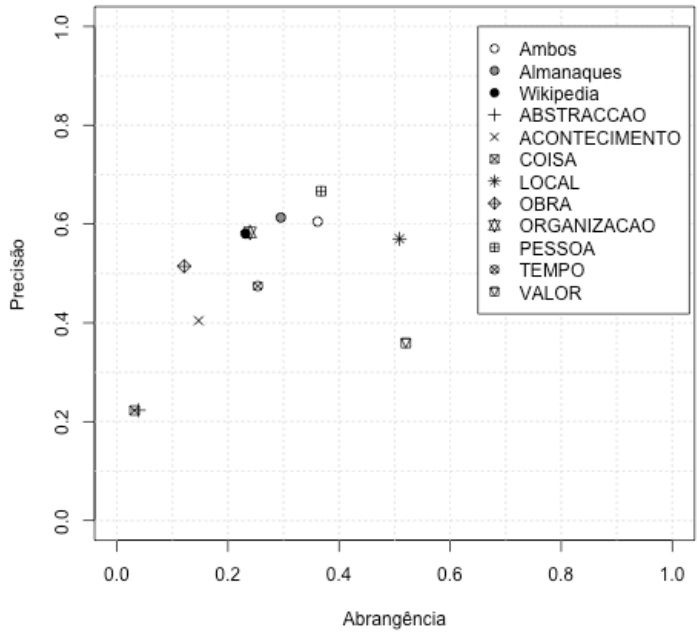

Figura 12.3: Distribuição de precisão e abrangência na tarefa de classificação das diversas categorias e corridas.

a categoria ABSTRACCAO (valor 1 no gráfico 12.4), tendo no entanto a sua abrangência e medida F sido a menor obtida por todos os sistemas participantes (igual a 0 no gráfico).

O gráfico permite observar que o REMMA obteve, de uma forma geral, resultados bastante precisos, estando sempre próximo do melhor sistema para cada uma das categorias. O contrário pode ser observado relativamente à abrangência.

A melhor classificação do REMMA, no cenário total, é relativa à corrida *Ambos*, tendo ficado na posição 10 em 29.

12.3.4 Comparação com o REMBRANDT

Após a realização do encontro do Segundo HAREM (Setembro de 2008) apercebemo-nos da existência de um sistema participante com uma abordagem bastante semelhante à do REMMA, o REMBRANDT (ver capítulo 11). Este sistema usa a informação relativa às categorias existentes na Wikipédia para a obtenção da classificação adequada a cada categoria do Segundo HAREM.

Relativamente a este sistema, comparando os resultados obtidos no cenário total e disponibilizados no capítulo 11, o REMMA obteve valores para a medida F inferiores aos ob-

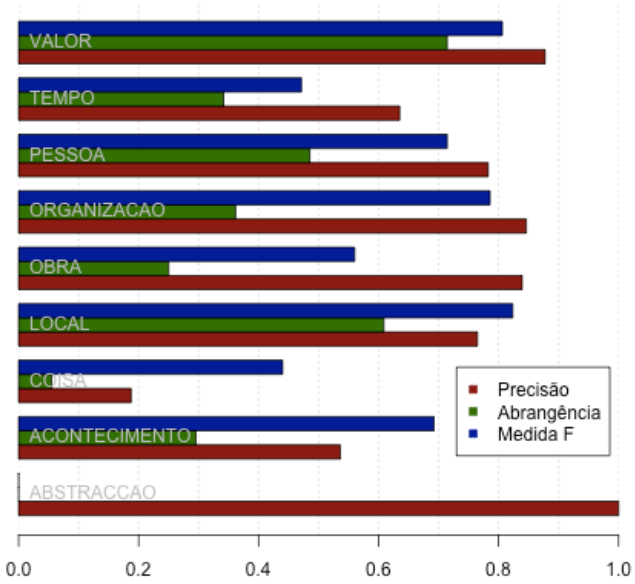

Figura 12.4: Comparação normalizada relativamente aos valores máximos e mínimos

tidos pelo REMBRANDT (cerca de 0,11 inferior, nas melhores corridas de ambos), tendo os sistemas, no entanto, precisões comparáveis.

No que diz respeito à análise individual de cada uma das categorias, observam-se melhores resultados do REMMA na classificação das entidades pertencentes à categoria ACONTECIMENTO (aproximadamente 0,02) e valores muito semelhantes para a categoria COISA, tendo para as restantes sido obtida pelo REMMA uma medida F inferior à do REMBRANDT.

12.4 Discussão

Os resultados obtidos pelo REMMA sugerem a utilidade da extracção de categorias semânticas da Wikipédia para a tarefa de REM, mesmo quando efectuada através de um método tão simples como o apresentado.

No entanto, estes indicam também que a utilização isolada da informação contida na Wikipédia, sem recurso a qualquer almanaque ou regra contextual, é uma solução com piores resultados, isto é, sugerem a existência de espaço para várias melhorias, como por exemplo, a necessidade de uma utilização mais abrangente das várias estruturas internas da Wikipédia.

Os resultados relativos às diferentes categorias avaliadas indicam uma imunidade das

categorias ABSTRACCAO e COISA ao uso da Wikipédia. Este resultado pode estar relacionado com uma maior ambiguidade, e consequente dificuldade, na categorização destas entidades. Outro factor que pode estar na origem destes resultados é o uso de um conjunto insuficiente de palavras na extração da classificação a partir da primeira frase do artigo. As categorias VALOR e TEMPO também não apresentam melhorias pelo uso da Wikipédia como fonte de conhecimento. Isto pode dever-se à introdução de categorias erradas aquando da utilização do anotador Wikipédia isoladamente, como por exemplo na análise de palavras relativas a estações do ano, meses, dias da semana ou unidades.

De uma forma geral, podemos afirmar que o REMMA é um sistema bastante preciso, tendo apresentado precisões competitivas relativamente aos demais sistemas. A comparação com o REMBRANDT em particular, indica precisões comparáveis, faltando ao REMMA abrangência. Isto deve-se, entre outros factores, à necessidade de criar mais regras e regras mais complexas para a identificação de EM candidatas, bem como métodos de resolução de conflitos e ambiguidades. Note-se, no entanto, que, no contexto da extracção de informação na área da medicina, importa a existência de um sistema preciso, capaz de anotar correctamente a informação existente, em oposição a um sistema que extraia muita informação com ruído.

Os resultados obtidos mostram, assim, existir espaço para várias melhorias. A tarefa de identificação do REMMA é realizada actualmente através de expressões regulares bastante simples, não contemplando expressões com algarismos. Uma melhoria nesta tarefa implicaria com certeza melhor abrangência. O facto de o anotador Wikipédia recolher informação de uma página apenas e só caso esta não seja um página de desambiguação pode evitar a introdução de ruído nos resultados, no entanto, a criação de métodos de resolução de conflitos entre entidades ou de desambiguação, bem como a utilização de outras estruturas internas disponíveis na Wikipédia, como é o caso das categorias, implicaria certamente melhorias significativas no REMMA.

De uma forma geral, a participação do REMMA no HAREM, embora direccionada a textos não especializados, permitiu perceber a utilidade de diferentes abordagens, acabando por indicar que a utilização de recursos e soluções semelhantes para a área em que nos concentramos, a medicina, é uma abordagem promissora, mesmo com a utilização de abordagens simples como a apresentada.

12.5 Conclusão e trabalho futuro

Para a participação no Segundo HAREM foi desenvolvido um sistema capaz de explorar fontes de conhecimento externas, como a Wikipédia, de modo a evitar a criação e a manutenção de almanaques de domínio geral de grande dimensão. A principal motivação para esta abordagem foi o carácter dispendioso desta tarefa, quer em termos de tempo, quer em termos dos recursos necessários. Foi desenvolvido um sistema composto por um conjunto de anotadores UIMA, capaz de usufruir de vários tipos de recursos, sejam estes almanaques simples de domínio geral, ou, categorias semânticas extraídas a partir da análise da primeira frase de um artigo da Wikipédia.

A utilização da Wikipédia demonstrou ser útil para a melhoria da classificação das entidades mencionadas, dando uma indicação clara da utilidade deste tipo de fontes de conhecimento e abrindo portas à procura e aplicação de soluções semelhantes a textos da área da medicina. Existem actualmente diversas wikis públicas e relativas a vários domí-

nios. O futuro do sistema REMMA passará, assim, pela utilização de recursos semelhantes relativos à área da medicina, de modo a melhorar a tarefa de extracção de informação que nos propomos realizar no âmbito do projecto MedAlert.

No entanto, ficou também claro neste trabalho a necessidade usar técnicas de desambiguação e de explorar outras estruturas internas disponibilizadas nas wikis públicas, como é o caso das categorias e das ligações internas na Wikipédia. Relativamente às páginas de desambiguação essa necessidade é mais evidente pelo facto de a Wikipédia ser uma enciclopédia em constante crescimento, o que implica um aumento constante do número de artigos e assuntos definidos e consequentemente, um aumento da ambiguidade das suas páginas. Uma interessante tarefa a realizar futuramente é o desenvolvimento de uma técnica de recuperação do título do artigo mais adequado ao contexto em questão, a partir de uma página de desambiguação.

A utilização do conteúdo da Wikipédia para a extracção de relações semânticas entre entidades é também uma interessante tarefa a realizar futuramente e uma área de grande interesse no âmbito do projecto MedAlert (Ferreira et al., 2008).

Agradecimentos

O projecto RTS foi financiado pelo programa "Aveiro Digital" da iniciativa "Portugal Digital" e pelo programa POSI do Governo Português. O projecto GERESmed é financiado pela Fundação para a Ciência e Tecnologia (GRID/GRI/81819/2006).