REMMA - Reconhecimento de Entidades Mencionadas do MedAlert

Liliana Ferreira e António Teixeira

Encontro do Segundo HAREM
Setembro 2008
Aveiro

Estrutura

ieeta

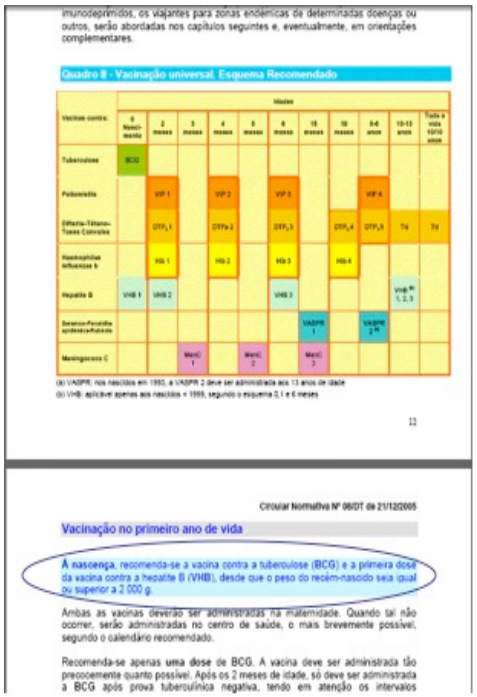
Motivação

Trabalho anterior em extracção de entidades ligadas à medicina:

a Extracção de Informação de Relatórios Médicos;

b MedAlert.

Exemplo: Plano Nacional de Vacinação



imunodeprimidos, os viajantes para zonas endémicas de determinadas doenças ou outros, serão abordadas nos capítulos seguintes e, eventualmente, em orientações complementares.

Quadro II - Vacinação universal. Esquema Recomendado

Circular Normativa Nº 06/DT de 21/12/2005

Vacinação no primeiro ano de vida

À nascença, recomenda-se a vacina contra a tuberculose (BCG) e a primeira dose da vacina contra a hepatite B (VHB), desde que o peso do recém-nascido seja igual ou superior a 2 000 g.

Ambas as vacinas deverão ser administradas na maternidade. Quando tal não ocorrer, serão administradas no centro de saúde, o mais brevemente possível, segundo o calendário recomendado.

Recomenda-se apenas uma dose de BCG. A vacina deve ser administrada tão precocemente quanto possível. Após os 2 meses de idade, só deve ser administrada a BCG após prova tuberculínica negativa, tendo em atenção os intervalos

⇒

À nascença, recomenda-se a vacina contra a tuberculose (BCG) e a primeira dose da vacina contra a hepatite B (VHB), desde que o peso do recém-nascido seja igual ou superior a 2 000 g.
Ambas as vacinas deverão ser administradas na maternidade.
Quando tal não ocorrer, serão administradas no centro de saúde, o mais brevemente possível, segundo o calendário recomendado.
Recomenda-se apenas uma dose de BCG.

REMMA - Reconhecimento de Entidades Mencionadas do MedAlert

- Desenvolvido em UIMA - *Unstructured Information Management Architecture*
- Principais recursos:

almanaques

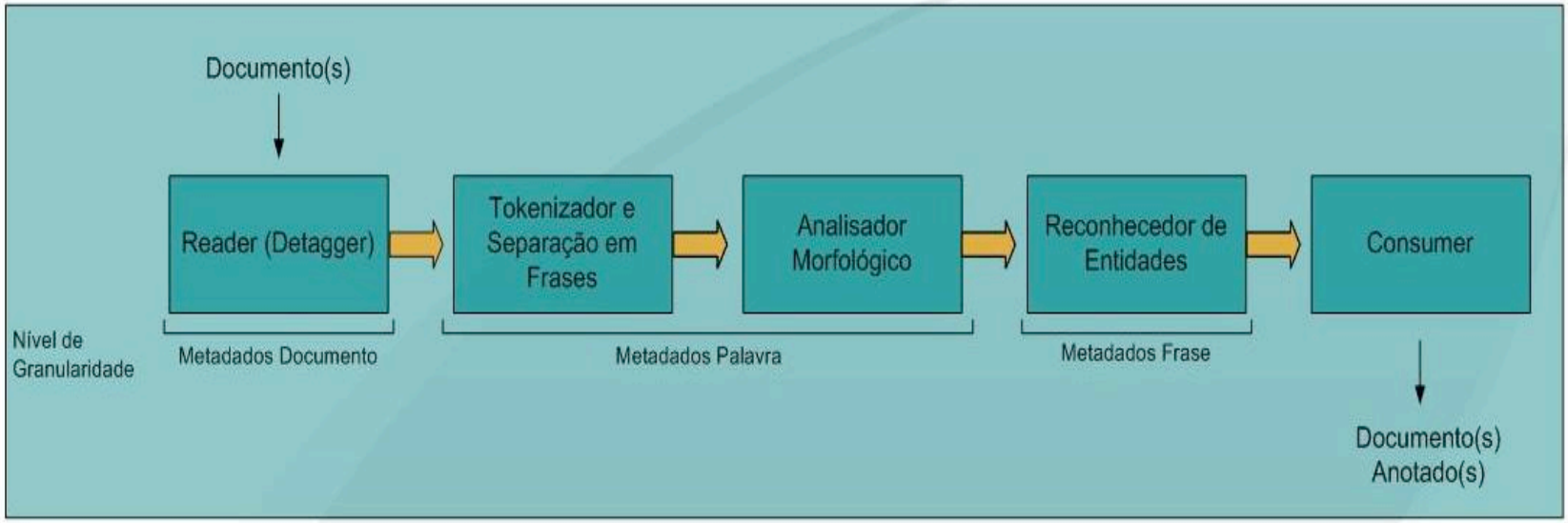
+

regras

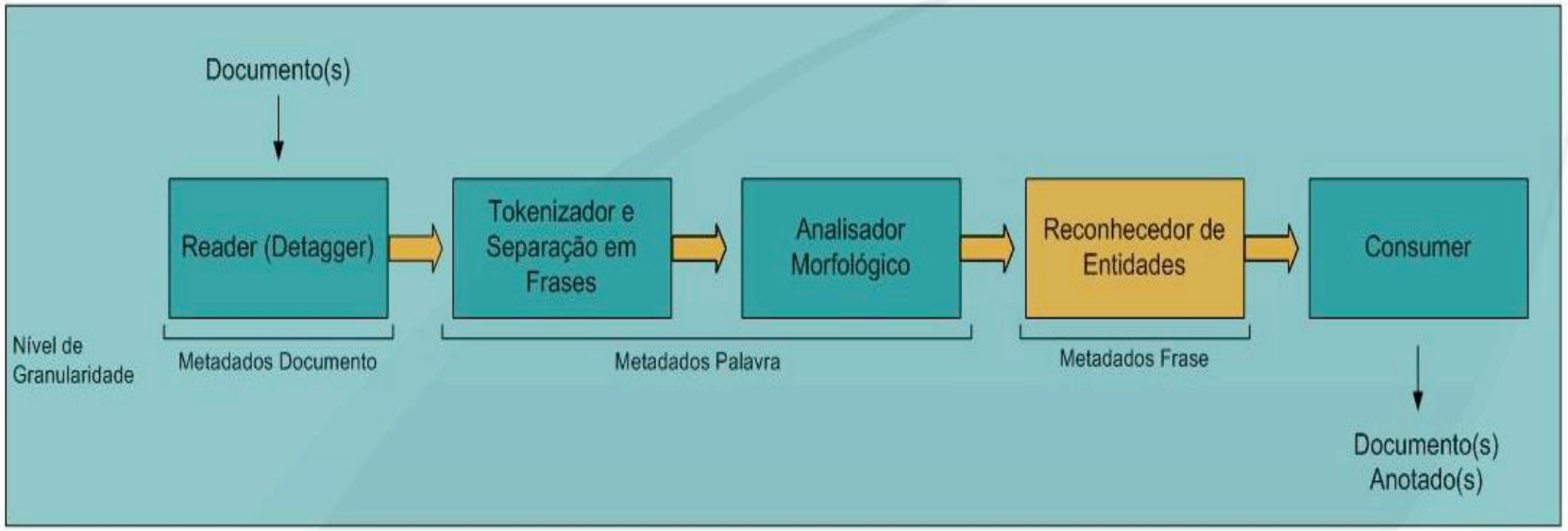
+

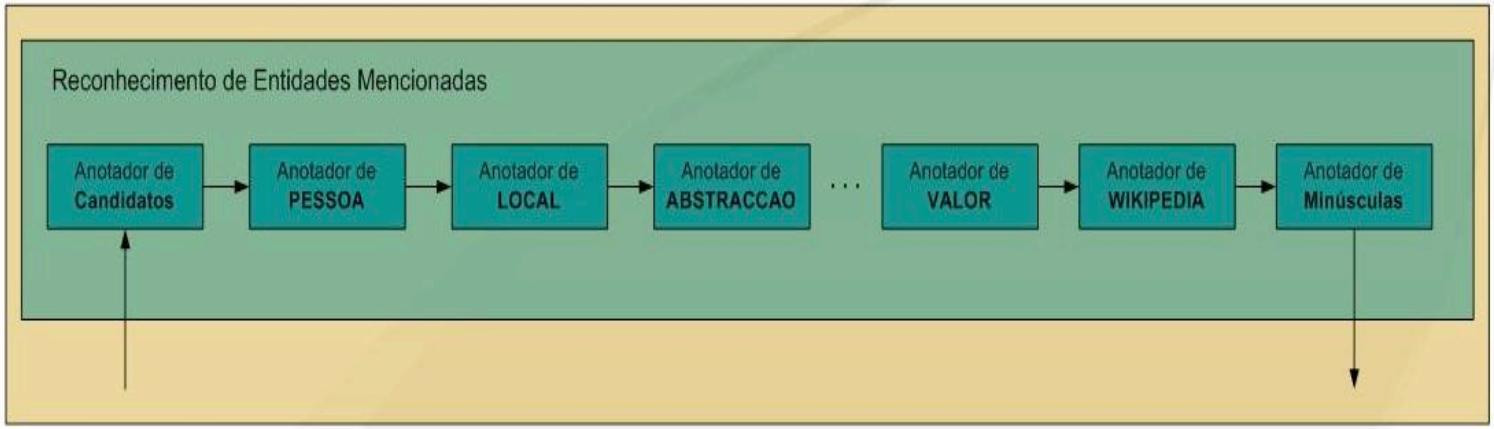
wikipedia

REMMA
Arquitectura



REMMA
Arquitectura





Problema:

- Almanaques ou dicionários de entidades são importantes em REM, no entanto, a sua construção e manutenção é extremamente demorada.

Solução REMMA:

- A wikipedia disponibiliza os conteúdos da base de dados para cada língua em XML e os ficheiros SQL necessários para construir a BD.

  `Dumps from any Wikimedia Foundation project:`
  `http://download.wikimedia.org/`

- Foi utilizada a versão portuguesa de Fevereiro de 2008, contendo cerca de 1 290 836 páginas.

Embora não exista uma convenção, o normal é começar o artigo por uma pequena frase descrevendo o que a entidade é:

Exemplo

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

Estrutura básica:

- Os artigos são identificados por um nome único - concatenação das palavras do título do artigo com "_"; *ex.: Universidade de Aveiro - Universidade_de_Aveiro*
- Obtemos o artigo correspondente à entidade, seguindo os redireccionamentos caso seja necessário.

Redirecção:

- Algumas páginas não têm artigo mas são direccionadas para outro artigo com um nome diferente.
- Estes mecanismo é denominado "redirecção"e é identificado por "#REDIRECT [[A B C]]".
- Nestes casos segue-se para o artigo com o título A_B_C.

Desambiguação:

- Alguns autores desenvolvem uma página de desambiguação para um artigo com nome ambíguo.
- Nós não usamos estas páginas!

Wikipedia como recurso externo
Método

Após a obtenção do artigo correspondente, extraímos a primeira frase da qual foi extraída a categoria:

Exemplo

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

$\Downarrow$

ORGANIZACAO

a Não usamos a secção *Categoria* dos artigos uma vez que um artigo pode ter mais que uma Categoria e muitas das apresentadas não são claramente hiperónimos da entidade.

REMMA

Participação

Cenário *Selectivo 4*: inclui todas as categorias e tipos previstos nas directivas mas não considera os subtipos associados, concretamente os subtipos de LOCAL e de TEMPO.

Resultados

1. Saída I: Regras + Almanaques + Wikipedia
   45 081 entidades em 52m06s
2. Saída II: Regras + Almanaques
   37 707 entidades em 32m24s
3. Saída III: Wikipedia
   29 949 entidades

Conclusões

Pretendemos demonstrar, com esta experiência, a utilidade de recursos deste tipo na classificação de entidades quando usados em conjunto com outros recursos como almanaques.

1. Explorar a *Wikipedia* como fonte de conhecimento externo:
   a. extracção a categoria a partir da primeira frase do artigo.

Trabalho Futuro:

1. Explorar a estrutura dos artigos:
   a. Categoria;
   b. Páginas de desambiguação;
   c. ...
2. Aplicar o método a outros recursos semelhantes na área da Medicina!

REMMA - Reconhecimento de Entidades Mencionadas do MedAlert

Liliana Ferreira e António Teixeira

Encontro do Segundo HAREM
Setembro 2008
Aveiro

Fim! Obrigado.