Extracção de Informação de Relatórios Médicos

Liliana Ferreira[1] César Telmo Oliveira[12] António Teixeira[1] João Paulo Silva Cunha[1]

[1] Instituto de Engenharia Electrónica e Telemática de Aveiro
Departamento de Electrónica, Telecomunicações e Informática
Universidade de Aveiro
3810-193 Aveiro, Portugal

[2] Hospital Infante D. Pedro
Avenida Artur Ravara
3814-501 Aveiro, Portugal
{lsferreira, ctelmo, ajst, jcunha}@ua.pt

Resumo

A utilização, cada vez mais frequente nos serviços de saúde nacionais, de sistemas de Registo Clínico Electrónico tem levado a um aumento significativo da informação disponível em formato electrónico. Embora muita desta informação exista, actualmente, numa forma estruturada, uma parte significativa encontra-se sob a forma de texto livre não estruturado. A necessidade de processar e gerir estas grandes quantidades de texto tem motivado o recente interesse em aproximações semânticas. Este artigo descreve o trabalho desenvolvido no âmbito do projecto MedAlert para a criação de um corpus anotado semanticamente e no desenvolvimento de um sistema de extracção automática de informação capaz de identificar entidades clínicas relevantes, bem como os seus relacionamentos. Para tal, o MedAlert possui actualmente um corpus de cerca de 48 000 textos médicos relativos a episódios de internamento ocorridos no Hospital Infante D. Pedro, em Aveiro. Um subconjunto do corpus foi seleccionado para a criação das directivas de anotação e anotação semântica manual e automática. O sistema de reconhecimento de entidades mencionadas REMMA foi usado numa primeira avaliação. Os primeiros resultados são apresentados indicando a necessidade de desenvolver directivas precisas para a anotação de textos médicos, de modo a melhorar a concordância entre anotadores.

1 Introdução

O acesso a informação clínica em instituições de saúde nacionais é feito, cada vez mais, através de variados sistemas de Registo Clínico Electrónico (RCE). Embora alguns relatórios médicos existam actualmente, nestes sistemas, numa forma estruturada, uma parte significativa é guardada ainda como texto livre não estruturado. Este é o caso dos relatórios relativos a episódios de internamento. Estes documentos contêm informação importante, não só para a manutenção do cuidado de saúde do doente, mas também de uso potencial em investigação. Descrevem, por exemplo, qual a medicação usada em cada tratamento, porque foi interrompida, quais os resultados de exames físicos e quais os problemas considerados relevantes na discussão com o paciente mas que nem sempre são considerados relevantes na codificação interna.

A necessidade de gerir este tipo de informação está a motivar aproximações semânticas, cujos principais objectivos são a redução de erros clínicos, a melhoria da eficiência, da segurança e da satisfação no serviço médico. Por exemplo, a informação contida nestes documentos poderia ser usada para assistir o clínico na formação de hipóteses, caso este pudesse obter respostas a questões relevantes, como por exemplo *Quantos pacientes com AVC isquémico agudo foram tratados com Enoxaparina e permaneceram sem outras complicações?* O tratamento individual de pacientes beneficiaria também, caso pudessem ser obtidos sumários concisos da história clínica do paciente ou se existisse acesso a histórias clínicas de pacientes com manifestações semelhantes reportadas em diversas ocasiões e localizações.

O MedAlert usa a tecnologia de extracção automática de informação nos dados disponibilizados no sistema de RCE em utilização no Hospital Infante D. Pedro em Aveiro, a Rede Telemática de Saúde (RTS) (Cunha et al., 2006).

Este artigo reporta a construção de uma colecção dourada para o projecto MedAlert, na qual os documentos clínicos são anotados com as suas múltiplas entidades e relacionamentos. Uma primeira avaliação do sistema de extracção au-

This work is licensed under a
Creative Commons Attribution 3.0 License

tomática de informação REMMA - *Reconhecimento de Entidades Mencionadas do MedAlert* é também apresentada.

A secção seguinte apresenta o projecto MedAlert e a sua motivação. A Secção 1.2 sumaria algum trabalho relacionado apresentado na literatura. Os recursos utilizados no MedAlert são apresentados na secção 2, onde é descrito o processo de selecção de documentos para a colecção dourada, o método de anotação usado e as respectivas entidades e relacionamentos. As fontes de conhecimento usadas na extracção automática de informação são descritas na secção 2.2. A secção 3 descreve o sistema REMMA e os primeiros resultados obtidos são discutidos na secção 4. O artigo termina na secção 5 com as conclusões e algumas sugestões de trabalho futuro.

1.1 MedAlert

Nos últimos anos tem sido realizado um investimento significativo em sistemas que permitam o acesso electrónico a informação clínica. Este tipo de acesso é cada vez mais uma realidade através de numerosos sistemas de RCE. No entanto, pouco tem sido feito na criação de sistemas que permitam a comunicação entre diferentes instituições médicas (Cunha et al., 2006). A Rede Telemática de Saúde (RTS)[1] tenta colmatar esta dificuldade através de uma infra-estrutura que permite a comunicação clínica entre os múltiplos serviços de saúde regional. Esta rede promove, assim, o acesso seguro a informação existente em vários serviços de saúde, a todos os profissionais credenciados. A RTS implementa um *Processo Clínico Electrónico Regional* resumido, que combina diversos documentos electrónicos existentes em todas as instituições que pertencem à rede, permitindo, assim, o acesso dos profissionais de saúde a informação como cartas de alta, resultados de exames e boletins de vacinação.

O MedAlert usa a informação disponibilizada pela RTS, 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 suporte e validação de decisões, melhorando, assim, o cuidado médico, com a redução de erros, melhoria de segurança e satisfação.

1.2 Trabalho relacionado

Várias aplicações de suporte à decisão clínica têm sido desenvolvidas recentemente, fazendo uso de técnicas de PLN e fontes de conhecimento como ontologias. Consequentemente, uma grande variedade de *corpora* anotados semanticamente e outras fontes de conhecimento médico foram desenvolvidas tendo em vista a investigação em extracção de informação biomédica. O *thesaurus Medical Subject Headings* (MeSH)[2] e o *Unified Medical Language System* (UMLS) (NLM, 2008), com as suas vertentes de *metathesaurus* e de rede semântica, são exemplos do esforço feito no sentido de facilitar o desenvolvimento de sistemas computacionais capazes de processar linguagem médica. Ambos são actualmente utilizados numa grande variedade de sistemas na catalogação, indexação e recolha de informação biomédica e de saúde.

Um esforço semelhante foi realizado no desenvolvimento do vocabulário trilingue DeCS - Descritores em Ciências da Saúde[3]. O DeCS foi desenvolvido a partir do MeSH com o objectivo de permitir o uso de terminologia comum para a pesquisa em três línguas, inglês, espanhol e português, proporcionando uma forma consistente e única para a recolha de informação médica. Os conceitos que compõem o DeCS são organizados numa estrutura hierárquica permitindo a execução de pesquisa em termos mais amplos ou mais específicos ou de todos os termos que pertençam a uma dada estrutura hierárquica.

2 Recursos

No desenvolvimento do sistema MedAlert são utilizados vários recursos, desde o *corpus* usado no desenvolvimento da colecção dourada MedAlert, até às várias fontes de conhecimento externo usadas na extracção automática de informação. Esta secção apresenta em mais detalhe estes recursos, começando por apresentar na Secção 2.1 o *corpus* MedAlert e o método usado na anotação semântica manual. A Secção 2.2 apresenta as fontes de conhecimento usadas no reconhecimento automático das entidades e relacionamentos definidos na anotação manual.

2.1 O *corpus* MedAlert

O *corpus* MedAlert é actualmente constituído por 48 229 textos relativos a episódios de internamento

[1] http://www.rtsaude.org

[2] http://www.nlm.nih.gov/mesh/

[3] http://decs.bvs.br/

ocorridos no Hospital Infante D.Pedro, em Aveiro. Estes relatórios incluem informação relativa a:

- Motivo de internamento;
- História clinica;
- Exame físico;
- Evolução;
- Terapêutica;
- Destino.

A Tabela 1 apresenta a distribuição de informação no *corpus*, em particular, a quantidade de documentos, frases e tokens existente para cada estrutura.

Os relatórios provêm do *Processo Clínico Electrónico Regional* implementado pela RTS, onde toda a informação confidencial relativa aos doentes e profissionais de saúde está já de uma forma estruturada e separada. Assim, os relatórios usados neste trabalho não contêm qualquer informação confidencial ou passível de identificação dos intervenientes no processo.

2.1.1 Colecção dourada MedAlert

A construção de uma colecção dourada MedAlert tem como objectivo servir três propósitos principais:

1. focar e clarificar os requisitos do sistema através da análise de dados anotados manualmente por peritos da área;
2. o desenvolvimento de um *gold standard* contra o qual os resultados da extracção automática de informação serão calculados;
3. o fornecimento de dados para o desenvolvimento do sistema: as regras de extracção podem deste modo ser criadas automaticamente ou manualmente, bem como podem ser desenvolvidos modelos estatísticos dos dados para a utilização de algoritmos de *machine learning*.

Dado o elevado custo da anotação manual, a ser realizada, neste caso, por pessoal médico especializado, a percentagem de relatórios a anotar teve de ser reduzida a um subconjunto relativamente pequeno de todo o corpus de 48 229 relatórios. Nesta fase inicial do processo e de modo a facilitar a introdução das directivas aos peritos, optou-se por focar nas estruturas Motivo de Internamento e História Clínica e num conjunto reduzido de documentos, embora no alcance dos objectivos finais do projecto seja necessária a existência de mais dados anotados manualmente e relativos a todas as estruturas dos relatórios.

Assim, optou-se pela utilização de um subconjunto de 120 relatórios, 20 para cada estrutura, tendo destes, 10 documentos sido usados no desenvolvimento das directivas de anotação e 10 na anotação manual.

Deste modo, a colecção dourada é constituída actualmente por 20 documentos anotados manualmente, relativos às estruturas Motivo de Internamento e História Clínica.

O restante artigo foca na anotação semântica e extracção automática de informação relativa aos relatórios de Motivo de Internamento.

2.1.2 Método de anotação

A construção de uma colecção dourada para o projecto MedAlert pressupõe a existência de um *corpus* de documentos médicos anotados semanticamente, quer com múltiplas entidades, quer com as suas relações.

De modo a garantir a qualidade da colecção dourada todos os documentos foram anotados pelo mesmo *standard* e foram desenvolvidas directivas especificas de modo a que as várias questões que surjam ao anotar os relatórios estejam devidamente esclarecidas. As directivas desenvolvidas pretendem, assim, garantir a consistência, descrevendo em detalhe o que deve e o que não deve ser anotado, respondendo a questões relevantes tais como, decidir se duas entidades estão relacionadas ou como lidar com correferência. As directivas apresentam também uma sequência de passos, uma receita, que os anotadores deverão seguir quando trabalham com os documentos, de modo a minimizar os erros de omissão. Deste modo, o desenvolvimento das directivas de anotação foi realizado através de um processo rigoroso e iterativo, criado de modo a garantir consistência (Roberts et al., 2007).

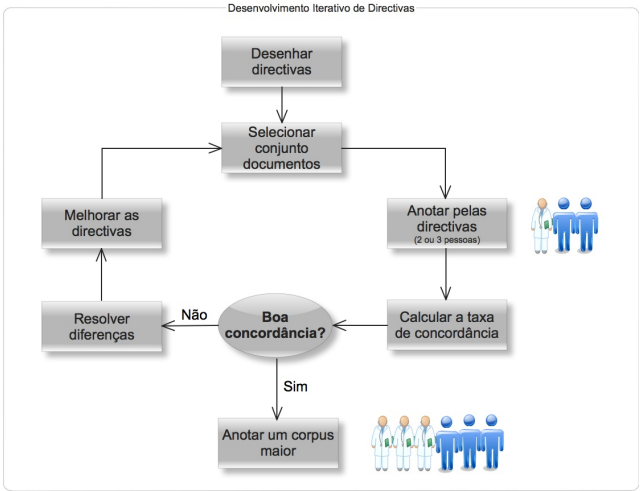

Figura 1: Processo Iterativo de anotação de relatórios.

Em detalhe o processo incluiu vários passos, apresentados na Figura 1, entre os quais se des-

Tabela 1: Relatórios MedAlert

Documento	Tokens	Frases	Textos
Motivo Internamento	104 833	11851	8 563
História Clinica	1 179 960	56 202	9 775
Exame Físico	414 558	37 499	7 071
Evolução	474 303	26 663	8 106
Terapêutica Efectuada	332 017	11 569	8 363
Destino	219 189	13 834	6 351
Total	**2 724 860**	**157 618**	**48 229**

taca:

1. Dupla anotação: um documento anotado por uma única pessoa pode reflectir vários problemas, como os valores ou erros frequentemente efectuados por um único anotador. A anotação dupla é uma forma comum de minimizar estes problemas, na qual cada documento é anotado independentemente por dois ou mais anotadores, e o conjunto de anotações comparado de modo a determinar a concordância.
2. Métricas de Concordância: o nível de concordância entre anotadores foi medido através do *índice de concordância inter-anotadores* (IAA):

$$IAA = \frac{\text{concordância}}{\text{concordância} + \text{não concordância}} \quad (1)$$

O índice de concordância foi calculado segundo um processo "relaxado", no qual as concordâncias parciais são contabilizadas como meia concordância. Os relacionamentos também foram avaliados usando IAA, tendo sido convencionado que apenas os relacionamentos envolvendo as entidades que todos os anotadores encontraram são contabilizados, permitindo, assim, isolar melhor a avaliação dos relacionamentos em relação à avaliação das entidades.

2.1.3 Entidades e Relacionamentos

Na definição da informação a anotar começou por definir-se os conceitos de entidade e relacionamento no contexto médico. Assim, *entidade* foi definida como algo real referido no texto, como por exemplo, a medicação mencionada, os exames realizados, etc. Os *relacionamentos* são então ligações entre entidades como o resultado de um exame ou a medicação indicada para uma patologia. A anotação também contemplou palavras que modificam marcações, tais como negação e caracterização. Duas ou mais marcações podem referir-se à mesma *entidade* real, e foram, neste caso, marcadas como *correferências*.

A Figura 2 apresenta alguns aspectos relevantes da anotação, tais como a marcação das entidades e dos seus relacionamentos.

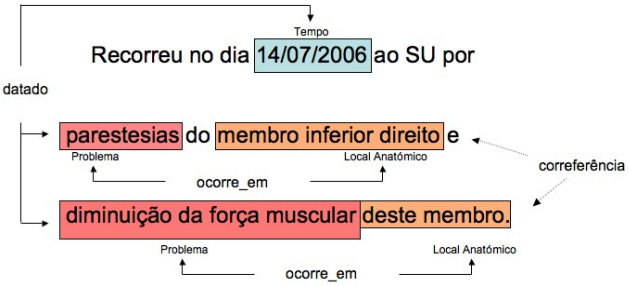

Figura 2: Exemplo ilustrativo de anotação.

A cada entidade e relacionamento foi atribuída uma *categoria*, tendo algumas sido classificadas também com o atributo *tipo*. No caso dos relatórios relativos ao Motivo de Internamento foram definidas as seguintes categorias:

- Problema - Sintomas, diagnósticos, complicações, condições e restantes problemas manifestados pelo doente;
- Local Anatómico - Estrutura ou localização anatómica, substância corporal ou função fisiológica, tipicamente a localização de um *Problema* ou *Exame*;
- Tempo - Expressões temporais, incluindo datas e tempos (absolutos e relativos), durações e frequências;
- Exame - Interacção entre o profissional de saúde e o doente com o objectivo de medir ou estudar algum aspecto do *Problema*;
- Resultado - Observação numérica ou qualitativa de um exame, excluindo referências a *Problemas*;
- Valor - Quantidades absolutas, relativas ou classificações;

- Caracterização - expressões que caracterizam outras entidades, como as pertencentes às categorias *Problema* e *Local Anatómico*;
- Negação - expressões que modificam outras entidades, neste caso negam, como por exemplo as entidades pertencentes à categoria *Problema* e *Resultado*.

Foram também definidos os seguintes relacionamentos:

- `inclui` - relação de *inclusão* entre entidades da mesma categoria, em particular aplicável às entidades das categorias *Problema, Local Anatómico* e *Exame*;
- `ocorre_em` - relação de *localização* entre um *Problema* ou *Exame* e o *Local Anatómico* em que é verificado;
- `datado` - relaciona as entidades *Exame*, *Problema* e *Resultado* com a sua indicação temporal (*Tempo*);
- `quantificado` - relaciona entidades quantificáveis, como as pertencentes às categorias *Resultado* ou *Problema* e o *Valor* que as caracteriza.
- `resulta` - relaciona um *Resultado* com o *Exame* que o produziu;
- `indica` - relaciona um *Problema* com o *Exame* que demonstrou a sua presença;
- `modificado` - relaciona um *Problema* ou *Resultado* com uma *Negação* ou *Caracterização*, bem como o *Local Anatómico* com a sua *Caracterização*, tal como a lateralidade: *direita, esquerda, bilateral* e sub-localização: *alto, baixo, extra, etc.*.

Alguns exemplos para cada uma das entidades e relacionamentos definidos, bem como os tipos atribuídos, são apresentados nas Tabelas 2 e 3.

De modo a facilitar o processo de anotação manual por parte dos especialistas, foram desenvolvidos esquemas de anotação para cada uma das estruturas dos documentos. O esquema de anotação relativo ao Motivo de Internamento é apresentado na figura 3, onde é possível visualizar cada uma das entidades definidas e a forma como estas se relacionam entre si.

2.1.4 Ferramentas de Anotação

De modo a realizar a anotação de uma forma consistente os esquemas de anotação foram modelados como ontologias Protégé-Frames[4] (Gennari et al., 2002). A anotação foi realizada usando o *plugin* Knowtator (Ogren, 2006) para Protégé. Este foi escolhido pelo facto de lidar com relacionamentos, após uma avaliação de outras ferramentas disponíveis (MMAX2[5], Wordfreak[6], Callisto[7]) e de arquitecturas de software de PLN como o GATE(Cunningham et al., 2002).

A Figura 4 apresenta a interface gráfica do Knowtator. No lado esquerdo da figura é possível visualizar o esquema de anotação criado para a anotação dos documentos do Motivo de Internamento. O quadro central e direito da figura apresenta um excerto de um relatório destacando a anotação da palavra `DPOC` como pertencente à classe `Diagnóstico` e o seu relacionamento de *inclusão* e *caracterização* com as palavras `insuficiência` e `agudizada`, respectivamente.

2.2 Fontes de conhecimento

O REMMA, sistema de Reconhecimento de Entidades Mencionadas do MedAlert, usa uma aproximação baseada em conhecimento de modo a detectar e classificar as expressões pertencentes às diversas categorias. Assim, várias fontes de conhecimento foram necessárias para a realização desta tarefa. Este é o caso da lista com cerca de 3 400 actos médicos e 1500 análises realizados no Hospital Infante D. Pedro, bem como da lista dos vários medicamentos disponíveis e comercializados em Portugal, com cerca de 12 800 entradas. Uma pequena lista com os nomes de problemas clínicos mais comuns, cerca de 200, foi também utilizada.

Apesar dos esforços realizados no sentido de obter o vocabulário biomédico DeCS - Descritores em Ciências da Saúde, tal não foi, até à data, possível. Assim, de modo a colmatar a falta de uma fonte de conhecimento especializada de grande abrangência, foi necessário recorrer a outras fontes de conhecimento não especializado como é o caso da Wikipédia. A secção seguinte faz uma pequena introdução à Wikipédia e à sua utilização em PLN.

2.2.1 Wikipedia

Recentemente, assistiu-se a um crescimento rápido e bem-sucedido da Wikipédia[8], uma enciclopédia electrónica livre e que está a ser construída por milhares de colaboradores em todo mundo. A Wikipédia tinha em Janeiro de 2009 mais de 2 700 000 artigos na versão inglesa e cerca de 454 000 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 que texto livre. A Wikipédia é actualizada diariamente, ou seja, novas entidades

[4] http://protege.standford.edu

[5] http://mmax.eml-research.de

[6] http://wordfreak.sourceforge.net

[7] http://callisto.mitre.org

[8] http://www.wikipedia.org

Tabela 2: Entidades MedAlert.

Categorias	Tipos	Exemplos
Problema	Sinal	*Prostração* marcada
	Sintoma	*Poliartralgias* MIs
	Diagnóstico	*Dpoc* agudizada
	Patologia	*Bronquite* Aguda
Local Anatómico		Hemorragia *digestiva* alta
Tempo	Tempo Calendário	Recorreu no dia *14/07/2006*...
	Duração	*Durante o internamento* ...
	Frequência	...a repetir a *cada meia hora*...
Exame	Físico	*Auscultação* pulmonar
	Analítico	...tendo sido realizada *biópsia*
	Imagiológico	cuja *EDA* revelou...
Resultado		Abdómen *sem alterações evidentes*
Valor		Lexotan *1,5mg*
Caracterização		Abdómen *sem* alterações evidentes
Negação		Acidente Vascular Cerebral *isquémico*

Tabela 3: Relacionamentos MedAlert.

Relacionamentos	Exemplos
inclui	[*arg1* dores] de garganta com [*arg2* tosse] e [*arg2* expectoração]
ocorre_em	[*arg1* dores] de [*arg2* garganta]
caracterizado_por	[*arg1* bronquite] [*arg2* aguda]
negado_por	[*arg2* sem] episódios prévios de [*arg1* convulsões]
datado_de	sem episódios [*arg2* prévios] de [*arg1* convulsões]
quantificado_por	[*arg1* febre] [*arg2* 40°C]
indica	realizou [*arg1* ecografia] abdominal que mostrou [*arg2* hepatoesplenomegalia] e [*arg2* esteatose] hepática
resulta	[*arg2* sem alterações] à [*arg1* auscultação]

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, Alfonseca, and Castells, 2006); (Santos et al., 2008); (Wu and Weld, 2007); (Zesch, Müller, and Gurevych, 2008)). A maior parte destes estudos concentram-se na extracção automática de almanaques da Wikipédia (Toral and Munoz, 2006) e na utilização da estrutura interna da Wikipédia para a desambigação de entidades mencionadas (Bunescu and Pasca, 2006). O REMMA baseia-se no método apresentado em (Kazama and Torisawa, 2007), onde se utiliza o sintagma nominal da primeira frase de um artigo 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. Por exemplo, o artigo Wikipédia sobre o *Acidente Vascular Cerebral* começa com a seguinte frase:

O Acidente Vascular Cerebral (AVC), ou Acidente Vascular Encefálico (AVE), vulgarmente chamado de"derrame cerebral", é caracterizado pela perda rápida de função neurológica, decorrente do entupimento ou rompimento de vasos sanguíneos cerebrais; é uma doença de início súbito, que pode ocorrer por dois motivos: isquemia ou hemorragia.

A extracção da palavra `doença` desta frase permite inferir a classificação a atribuir à entidade *Acidente Vascular Cerebral*. O método utilizado na obtenção destas classificações é descrito em detalhe na secção 3.

A Wikipédia disponibiliza o todo conteúdo para cada uma das diferentes línguas, em formato XML, bem como as ferramentas necessárias para a sua conversão para SQL, formato utilizado pelo REMMA na tarefa de classificação de entidades[9].

[9] O esquema completo da base de dados pode ser consultado em `http://www.mediawiki.org/wiki/Manual:Database_layout`

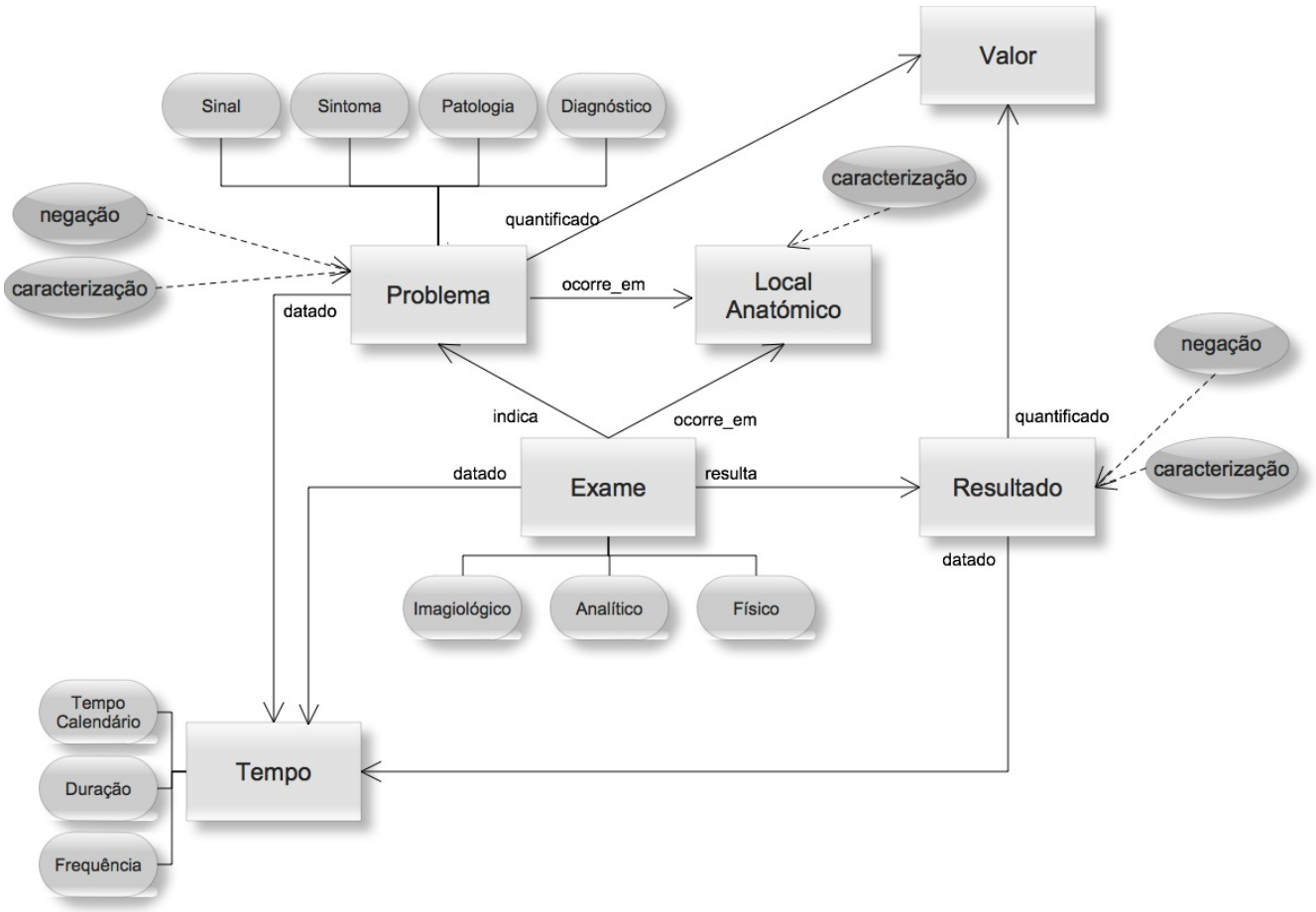

Figura 3: Esquema de anotação do Motivo de Internamento.

3 O sistema REMMA

O sistema REMMA foi inicialmente desenvolvido tendo em vista a participação no Segundo HAREM (Mota and Santos, 2008), uma avaliação conjunta na área do reconhecimento de entidades mencionadas em português, realizada em Abril de 2008. Para este evento o REMMA tinha como objectivo o reconhecimento de entidades mencionadas em textos de domínio geral, principalmente noticiosos (Ferreira, Teixeira, and Cunha, 2008). Para a extracção de informação de textos médicos, especificamente relativos a motivos de episódios de internamento hospitalar, várias adaptações foram realizadas. A secção seguinte descreve a arquitectura e os métodos usados para a identificação e classificação semântica das entidades e relacionamentos destes relatórios.

3.1 Arquitectura

Uma característica do sistema é a sua integração na plataforma UIMA. O UIMA, *Unstructured Information Management Architecture* (Ferrucci and 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 na modelação em vez de na programação.

O UIMA usa uma Estrutura de Análise Comum (CAS, *Common Analysis Structure*) 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. O CAS é partilhado entre os diversos anotadores que processam a colecção de objectos, passando de um anotador para seguinte no processo.

A arquitectura do REMMA está apresentada na Figura 5.

O REMMA começa por ler os documentos, um por um, e guardar os respectivos metadados.

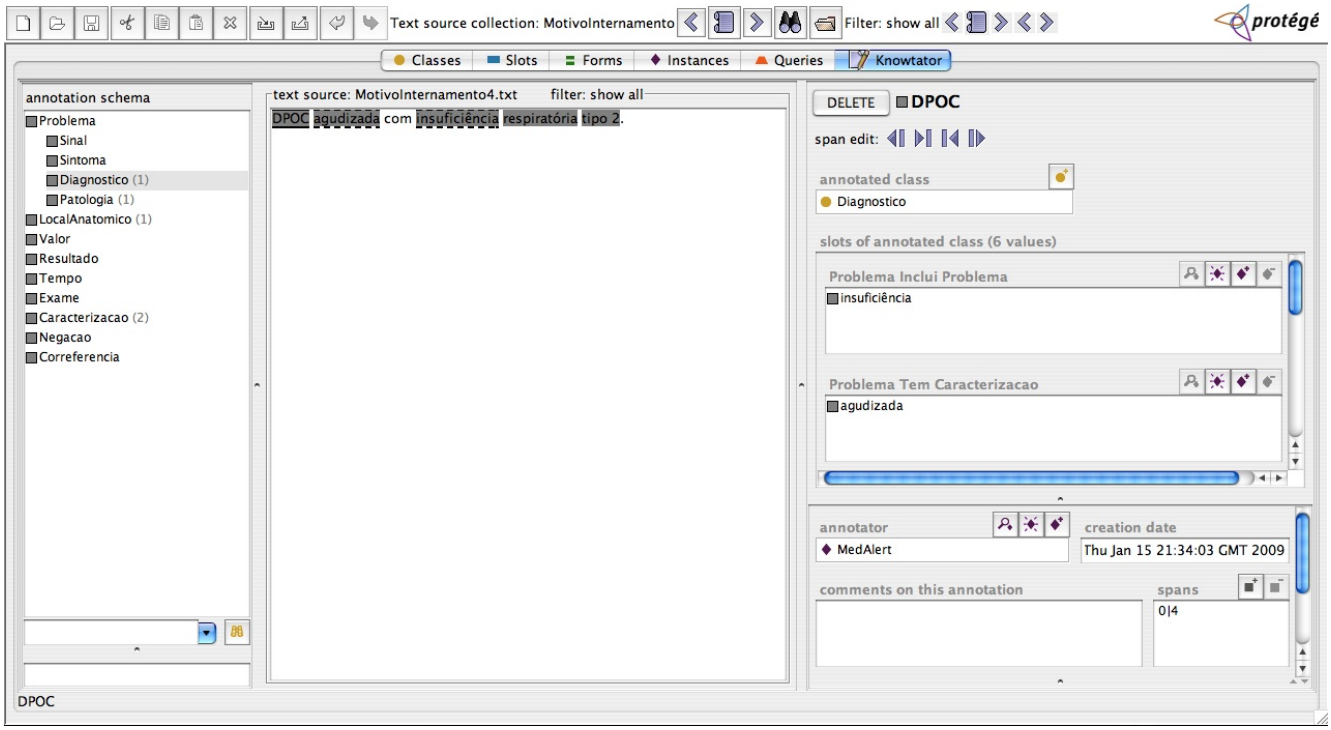

Figura 4: Motivo de Internamento no knowtator.

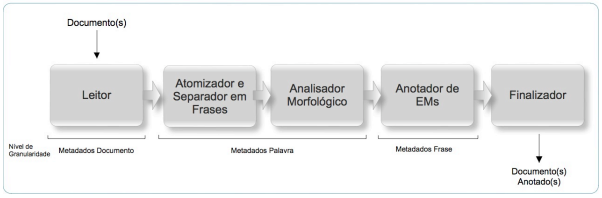

Figura 5: Arquitectura do REMMA

Os textos são posteriormente divididos em frases e tokens com a ajuda das ferramentas de pré-processamento disponíveis no UIMA. O analisador TreeTagger (Schmid, 1995) foi usado na obtenção das categorias morfossintáticas.

As anotações geradas por estas ferramentas são armazenadas no CAS e usadas nos diversos anotadores que constituem o módulo de extracção de informação. A Figura 6 apresenta a sequência de anotadores utilizados na identificação e classificação das entidades e relacionamentos. Estes anotadores são apresentados em mais detalhe nas secções seguintes.

O primeiro anotador a ser invocado é o Anotador de Candidatos que identifica excertos de frases com mais possibilidade de conterem entidades. As expressões candidatas são todos os conjuntos de palavras separadas por termos de ligação como as preposições *com* ou *por*, ou por pontuação. Estas expressões candidatas são posteriormente analisadas pelos anotadores de classificação.

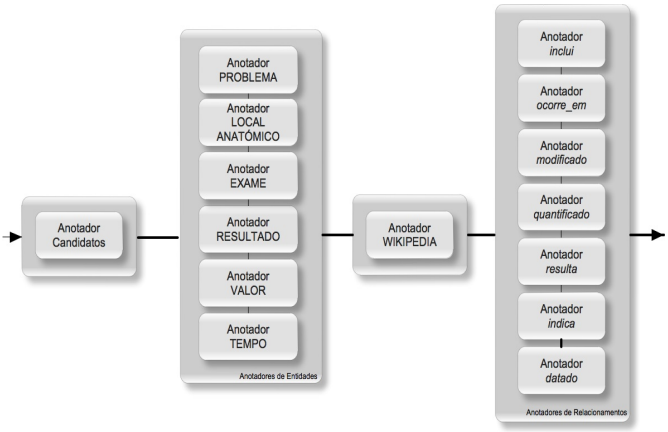

Figura 6: Anotadores do REMMA

O REMMA foi desenvolvido de modo a contemplar duas abordagens de classificação distintas. A primeira baseia-se em almanaques e regras muito simples, apresentada na secção 3.1.1 e a segunda é realizada com base na informação extraída da Wikipédia. Esta ultima é descrita em mais detalhe na secção 3.1.2. Os anotadores desenvolvidos para a identificação e classificação de relacionamentos são descritos na secção 3.1.3

3.1.1 Classificação semântica 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 das fontes de conhecimento externas descritas na Secção 2.2.
As regras utilizadas foram criadas manualmente e baseiam-se, não só no contexto em que a expressão é referida, mas também na existência de palavras com prefixos ou sufixos indicativos de classificação semântica. Por exemplo, na identificação e classificação de termos pertencentes à classe semântica *Problema* foram procuradas expressões começadas por `Síndrome de` ou `Insuficiência`, bem como palavras começadas por `hiper, hipo, hemo` ou terminadas em `patia, algia, ismo, ose, oma`.

Os anotadores que usam a informação contida nestes almanaques e regras começam por dividir a expressão candidata nos seus vários termos e atribuem categoria semântica caso algum dos termos da expressão exista nas listas usadas. Quando esta anotação não é conseguida, aplicam na expressão candidata as regras contextuais desenvolvidas para a classe semântica em análise.

3.1.2 Classificação semântica com recurso à Wikipédia

Na tarefa de classificação semântica 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.

O Anotador Wikipédia foi desenvolvido de modo a encontrar uma entidade na Wikipédia correspondente à identificada nos textos em análise. Deste modo, cada um dos termos existentes nas entidades candidatas identificadas é convertido num nome de entidade Wikipédia através da concatenação dos vários termos da expressão com o caracter "_". Por exemplo, a expressão *Acidente Vascular Cerebral* é convertida em `Acidente_vascular_cerebral` e o artigo correspondente recuperado.

Embora não exista uma regra de formatação estrita, é normal que os artigos Wikipédia comecem com uma pequena frase que define a entidade descrita no artigo. Por exemplo, como foi visto anteriormente o artigo com o título `Acidente_vascular_cerebral` ou `AVC` contém a frase:

O Acidente Vascular Cerebral (AVC) ... é uma doença de início súbito, que pode ocorrer por dois motivos: isquemia ou hemorragia

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, a palavra *doença*.

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. 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 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 4.

Tabela 4: Exemplos e quantidade de palavras-chave usadas na extracção de uma categoria semântica da primeira frase de um artigo.

Categoria	Exemplos

Problema (N=13)	`doença`
	`trauma`
	`sintoma`
	. . .

Local Anatómico (N=6)	`corpo humano`
	`órgão`
	`sistema`
	. . .

Exame (N=5)	`exame`
	`método de diagnóstico`
	`meio complementar de diagnóstico`
	. . .

3.1.3 Identificação e classificação de relacionamentos entre entidades

O anotador de relacionamentos do REMMA usa ainda um método muito simples e inicial para a detecção de relacionamentos entre entidades. Este usa a informação relativa às várias entidades identificadas nos passos anteriores, em conjunto com os termos de ligação usados pelo anotador de candidatos na identificação dos termos candidatos.

Especificamente, este anotador analisa as entidades identificadas e classificadas em cada uma das expressões candidatas e determina a presença na mesma expressão candidata de entidades pertencentes a categorias relacionáveis, por exemplo, caso uma expressão candidata contenha entidades pertencentes às categorias *Problema* e *Caracterização*, o relacionamento `modificado` é marcado entre estas entidades.

Um método particular é utilizado na identificação dos relacionamento de *inclusão*. Neste

caso, todas as sequências de expressões candidatas ligadas pela preposição *com* são analisadas. Caso ambas contenham pelo menos uma entidade pertencente às categorias *Problema* ou *Exame*, estas são marcadas como relacionadas.

Após a anotação das entidades identificadas pelos vários métodos descritos, um último anotador é chamado, o Finalizador. Este anotador analisa o 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). Um exemplo da saída gerada por este anotador é apresentado de seguida. No exemplo, as entidades identificadas são marcadas com a etiqueta equivalente ao nome da entidade, sendo ainda atribuída uma identificação única, `ID`, usada na marcação dos relacionamentos entre entidades.

```
<PROBLEMA ID='p1'
    REL='c6' TIPOREL='caracterizado'
    REL='p20' TIPOREL='inclui'>
      DPOC
</PROBLEMA>
<CARACTERIZACAO ID='c6'>
      agudizada
</CARACTERIZACAO>
       com
<PROBLEMA ID='p20'
   REL='l34' TIPOREL='ocorre_em'
   REL='c47' TIPOREL='caracterizado'>
      insuficiência
</PROBLEMA>
<LOCAL ID='l34'>
      respiratória
</LOCAL>
 <CARACTERIZACAO ID='c47'>
      tipo 2
</CARACTERIZACAO> .
```

4 Resultados

Ao longo do processo de criação da colecção dourada MedAlert diversas avaliações foram efectuadas. A secção 4.1 apresenta os resultados obtidos no processo de definição das directivas de anotação e posterior anotação manual. A secção 4.2 concentra-se nos resultados obtidos na tarefa de reconhecimento automático de entidades e dos relacionamentos entre estas.

4.1 Anotação Manual

Na construção das directivas finais para a anotação da colecção dourada foi obtido um nível de concordância (IAA) de 100%, quer na anotação manual de entidades, quer na anotação de relacionamentos entre estas.

A anotação manual da colecção dourada foi realizada por dois anotadores que seguiram os vários passos e conceitos descritos nas directivas desenvolvidas. Um dos anotadores possui conhecimento médico especializado, mas não tem conhecimentos de processamento de linguagem natural, enquanto que o outro anotador não possui qualquer conhecimento médico especializado, mas tem alguma experiência na anotação de colecções de texto médico.

O nível de concordância inter-anotadores obtido na anotação manual das entidades e seus relacionamentos é apresentado nas tabelas 5 e 6, respectivamente, onde se verifica um IAA de 80% para a anotação de entidades e de 66% na anotação de relacionamentos. Relembramos que apenas os relacionamentos que ambos os anotadores encontraram foram contabilizados.

Estes resultados demonstram claramente a dificuldade, não só na definição de directivas claras em áreas tão especializadas como a medicina, mas também em conseguir que os anotadores sigam as directivas de uma forma consistente. Foram verificados vários problemas como a não concordância em limites de entidades, a inclusão ou não de preposições nas entidades, a dificuldades em separar o conceito de caracterização ou negação, ou mesmo os conceitos de caracterização e local anatómico. Na anotação de relacionamentos verificou-se uma dificuldade acrescida na definição de quais as entidades envolvidas no relacionamento. Por exemplo, qual a entidade caracterizada ou qual a entidade que inclui outra entidade.

4.2 Extracção de Informação

Os resultados obtidos na tarefa de reconhecimento de entidades e relacionamentos são sumariados nas tabelas 7 e 8, respectivamente. As linhas apresentam o número de entidades e relacionamentos correctamente anotados pelo sistema, parcialmente correctos, falsos positivos e as entidades e relacionamentos que o sistema não foi capaz de identificar. Os resultados em termos de Precisão, Abrangência e Medida F estão nas linhas finais da tabela.

Uma precisão de 100% foi obtida para as entidades `LOCAL_ANATOMICO` e `TEMPO`, bem como para os diversos relacionamentos definidos, excepto para o relacionamento `datado`. Estes resultados permitem afirmar que o REMMA, embora esteja ainda numa fase inicial de adaptação à área médica e usando ainda métodos muito simples, é um sistema bastante preciso. Note-se que no contexto da extracção de informação na área da medicina, importa a existência de um sistema preciso, capaz de anotar a informação existente, em oposição a um sistema que extraia muita informação incor-

Tabela 5: Índice de concordância inter-anotadores na anotação manual das entidades

	PROBLEMA	LOCAL_ANATOMICO	CARACTERIZACAO	TEMPO	**Total**
Concordância	20	9	9	0	**38**
Concordância parcial	4	2	0	1	**7**
Não concordância	1	3	4	2	**10**
IAA	**0,96**	**0,89**	**0,69**	**0,20**	**0,80**

Tabela 6: Índice de concordância inter-anotadores na anotação manual dos relacionamentos

	inclui	ocorre_em	caracterizado	datado	**Total**
Concordância	3	11	7	0	**21**
Concordância parcial	0	0	0	0	**0**
Não concordância	2	3	5	1	**11**
IAA	**0,60**	**0,78**	**0,58**	**0,00**	**0,66**

recta ou com ruído.

É de notar a presença em alguns relatórios de problemas na escrita de algumas palavras, situação comum na escrita deste tipo de relatórios descritivos realizados em simultâneo ou imediatamente após a observação do paciente. Um exemplo comum deste tipo de problema é a escrita da palavra `disgestiva` em vez de `digestiva` dificultando a procura do seu significado nas fontes de conhecimento usadas pelo REMMA.

5 Conclusões

Para a extracção automática de informação de relatórios médicos é indispensável a existência de um *corpus* anotado semanticamente, quer com múltiplas entidades, quer com as suas relações. Para tal, foi apresentada uma metodologia para a anotação manual de uma colecção dourada de relatórios médicos de episódios de internamento hospitalar. Esta colecção dourada pretende auxiliar o processo de extracção de informação e sua avaliação. Os resultados iniciais mostram a importância da criação de directivas claras e precisas de modo a atingir bons valores de concordância entre anotadores, bem como a necessidade de coordenação e motivação entre anotadores.

Para a extracção de informação foi utilizado o sistema REMMA, um sistema composto por um conjunto de anotadores UIMA, capaz de usufruir de vários tipos de recursos, sejam estes almanaques especializados, ou, categorias semânticas extraídas a partir da análise da primeira frase de um artigo da Wikipédia. Apesar estar ainda em fase inicial, o REMMA apresenta resultados consistentes com um sistema bastante preciso, característica importante em sistemas de apoio à decisão médica.

5.1 Trabalho Futuro

O projecto MedAlert pretende actuar como um sistema de apoio à decisão clínica, capaz por exemplo, de inferir de uma forma automática dúvidas suscitadas pelas decisões médicas através da análise de relatórios médicos e de textos contendo directivas médicas. Assim, o aumento do conjunto de textos anotados semanticamente, textos estes pertencentes a todas as fases relativas ao processo de internamento hospital, é crucial no desenvolvimento de um sistema útil. De modo a melhorar a qualidade da anotação, é também essencial o aumento do leque de anotadores especializados.

A utilização da Wikipédia no REMMA foi útil para a melhoria da classificação das entidades mencionadas, dando uma indicação clara da utilidade deste tipo de fontes de conhecimento. Existem actualmente diversas wikis públicas e relativas a vários domínios. O futuro do sistema REMMA poderá passar, assim, pela utilização de recursos semelhantes relativos à área da medicina, de modo a melhorar a tarefa de extracção de informação. No entanto, o acesso e utilização de fontes de conhecimento especializadas, em particular o acesso ao vocabulário biomédico DeCS, é uma das tarefas prementes no âmbito do projecto MedAlert. Este tipo de informação segue uma estrutura bem definida e aceite internacionalmente, pelo que permite a estandardização das regras a serem aplicadas em sistemas como o MedAlert.

A natureza descritiva e expontânea dos relatórios médicos analisados, escritos em contexto de consulta hospitalar, leva à existência de vários erros ortográficos. Esta situação é mais grave quando se utilizam de sistemas de extracção de informação baseados em fontes de conhecimento, como é o caso do REMMA. Este problema ficou demonstrado nos resultados obtidos. Assim, a utilização e adaptação de um sistema de correcção ortográfica à área da medicina é um dos próximos passos do projecto MedAlert.

Agradecimentos

O projecto RTS foi financiado pelo programa "Aveiro Digital" da iniciativa "Portugal Digital"

Tabela 7: Resultados na tarefa de reconhecimento de entidades

	PROBLEMA	LOCAL_ANATOMICO	CARACTERIZACAO	TEMPO	**Total**
Saídas correctas	22	13	9	1	45
Parcialmente correctas	3	0	1	0	4
Falsos positivos	0	0	0	0	0
Em falta	3	3	2	0	8
Total	**28**	**16**	**12**	**1**	**57**
Precisão	0,88	1,00	0,90	1,00	0,92
Abrangência	0,89	0,81	0,83	1,00	0,86
Medida F	**0,88**	**0,89**	**0,86**	**1,00**	**0,89**

Tabela 8: Resultados na tarefa de reconhecimento de relacionamentos

	inclui	ocorre_em	caracterizado	datado	**Total**
Saídas correctas	4	11	7	0	22
Parcialmente correctas	0	0	0	0	0
Falsos positivos	0	0	0	0	0
Em falta	1	3	3	2	9
Total	**5**	**14**	**10**	**2**	**31**
Precisão	1,00	1,00	1,00	0,00	1,00
Abrangência	0,80	0,78	0,70	0,00	0,71
Medida F	**0,89**	**0,88**	**0,82**	**0**	**0,82**

e pelo programa POSI do Governo Português.

References

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *In 6th Int'l Semantic Web Conference, Busan, Korea*, pages 11–15. Springer.

Bunescu, Razvan and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Abril.

Cunha, João Paulo Silva, Isabel Cruz, Ilídio Oliveira, António Sousa Pereira, César Telmo Costa, Ana Margarida Oliveira, and Amândio Pereira. 2006. The RTS project: Promoting secure and effective clinical telematic communication within the Aveiro region. In *Em eHealth 2006 High Level Conference*, pages 1–10, Maio.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Julho.

Ferreira, Liliana, António Teixeira, and João Paulo Silva Cunha. 2008. REMMA- Reconhecimento de Entidades Mencionadas do MedAlert. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, Aveiro, Portugal, 7 de Setembro.

Ferrucci, David and Adam Lally. 2004. UIMA an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

Gennari, John H., Mark A. Musen, Ray W. Fergerson, William E. Grosso, Monica Crubézy, Henrik Eriksson, Natalya F. Noy, and Samson W. Tu. 2002. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. *International Journal of Human-Computer Studies*, 58:89–123.

Kazama, Jun'ichi and Kentaro Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707, June.

Mota, Cristina and Diana Santos, editors. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, Aveiro, Portugal, 7 de Setembro.

NLM, editor. 2008. *UMLS Knowledge Sources*. National Library of Medicine, Novembro.

Ogren, Philip. 2006. knowtator: A plug-in for creating training and evaluation data sets for biomedical natural language systems. In *Proceedings of the 9th INternational Protégé Conference*, pages 73–76, Stanford, California.

Roberts, A., R. Gaizauskas, M. Hepple, N. Davis, G. Demetriou, Y. Guo, J. Kola, I. Roberts, A. Setzer, A. Tapuria, and B. Wheeldin. 2007. The CLEF Corpus: Semantic Annotation of Clinical Text. In J. M. Teich, J. Suermondt, and G. Hripcsak, editors, *American Medical Informatics Association 2007 Proceedings. Biomedical and Health Informatics: From Foundations to Applications to Policy*, pages 625–629, Chicago, IL, USA, November. American Medical Informatics Association.

Ruiz-Casado, Maria, Enrique Alfonseca, and Pablo Castells. 2006. From wikipedia to semantic relationships: a semi-automated annotation approach. In *1st Workshop on Semantic Wikis: From Wiki to Semantics, at the 3rd European Semantic Web Conference (ESWC 2006)*, Junho.

Santos, Diana, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. 2008. Getting geographical answers from Wikipedia: the GIKIP pilot at CLEF. In *Working notes for the Cross Language Evaluation Forum, CLEF'2008*, 17–19 Setembro.

Schmid, Helmut. 1995. TreeTagger, a language independent part-of-speech tagger. *Institut fur Maschinelle Sprachverarbeitung, Universidade de Estugarda.*

Toral, Antonio and Rafael Munoz. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition using wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Abril.

Voss, Jakob. 2005. Measuring Wikipedia. In *10th International Conference of the International Society for Scientometrics and Informatics*, pages 221–231, Julho.

Wu, Fei and Daniel S. Weld. 2007. Autonomously semantifying wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50, New York, NY, USA. ACM.

Zesch, Torsten, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Maio.