Extracção de Informação de Relatórios Médicos

Liliana Ferreira, António Teixeira, João Paulo Cunha

Abstract – **This paper presents the first steps given to develop an information extraction system for portuguese texts. The system intends the extraction of information from medical reports and is based on the GATE system developed in the University of Sheffield. We present the changes made to this system in order to adapt it to the information extraction in Portuguese and some examples of results already gotten.**

Resumo – **Neste artigo apresentam-se os primeiros passos dados no sentido de desenvolver um protótipo de um sistema para a extracção de informação em Português. O sistema tem como domínio de aplicação relatórios médicos da área da neurofisiologia e baseia-se no sistema GATE desenvolvido na Universidade de Sheffield. As alterações efectuadas a este sistema com o intuito de o adaptar à extracção de informação em português são apresentadas, bem como alguns exemplos de resultados já obtidos.**

Palavras chave – **Extracção de Informação, Processamento de Linguagem Natural, Recuperação de Informação, *Text Mining***

I. Introdução

A Extracção de Informação é um dos campos de uma área de estudos mais abrangente: o Processamento de Linguagem Natural, que estuda os idiomas humanos a partir de uma perspectiva computacional. Outros campos desta área são, por exemplo, a Extracção de Conhecimento de Texto (*Text Mining*), que tem como objectivo a descoberta, reconhecimento e derivação de nova informação a partir de um grande conjunto de textos, e a Recuperação de Informação (*Information Retrieval*), cujo objectivo passa pela obtenção de informação relevante a partir de um amplo conjunto de textos, sendo as suas técnicas tipicamente utilizadas para obter documentos relevantes a partir de um conjunto de vários tipos de documentos, entre outras.

A extracção de informação (doravante IE do inglês *Information Extraction*) automática de textos envolve decidir se um texto é relevante para um dado domínio e caso seja, extrair um conjunto de factos desse texto. No entanto, a maior parte dos sistemas de IE foram desenvolvidos para textos escritos na língua inglesa. Actualmente, a IE em inglês está muito próxima do desempenho de especialistas humanos.

Um tipo de informação que abunda nos ambientes hospitalares é a informação escrita, cada vez mais em formato digital, e a informação transmitida oralmente entre vários intervenientes nos processos clínicos. Num cenário em que vingue a utilização sistemática de sistemas de transcrição de relatórios de uma forma automática, será necessário que os sistemas tenham cada vez mais capacidades de associar significado às palavras e frases com que lidam. Mesmo noutros ambientes, como a web, assiste-se a evoluções no sentido de uma *semantic web* em que a informação terá *tags* facilitando a procura por conceitos e não pelas actuais palavras.

Deste modo, o desenvolvimento de sistemas que obtenham informação existente em relatórios médicos de uma forma automática tornaria acessível a grande quantidade de informação deste tipo existente em ambientes hospitalares.

II. Extracção de Informação

A IE é uma tecnologia que transforma dados não-estruturados de documentos, em informações explicitas; isola partes relevantes do texto, extrai informação dessas partes e transforma-as em informações mais digeridas e melhor analisadas. Além disto, permite também formatar as informações recolhidas nos textos construindo padrões de saída (por exemplo, bancos de dados estruturados ou frases em linguagem natural).

O objectivo da IE direcciona-se para a necessidade de recolher informação produzindo dados estruturados a partir de um número indefinido de textos, permitindo o desenvolvimento de processos de base de conhecimento fundamentado, ou seja, modelos específicos de ocorrências, entidades ou relações.

No entanto, do ponto de vista do processamento de linguagem natural (NLP do inglês *Natural Language Processing*), a IE é atractiva pois, as suas tarefas estão bem definidas. Para além disto, a IE utiliza textos reais e coloca problemas de NLP difíceis e interessantes. A *performance* da IE pode ser comparada à *performance* humana na mesma tarefa.

A. Dificuldades que se colocam à IE

A.1 Portabilidade

Uma das barreiras que se apresentam à IE é o custo de adaptar um sistema de extracção a um novo "cenário". Em geral cada aplicação de IE envolve um cenário diferente e se a implementação deste novo cenário exigir meses de trabalho e de intervenção dos *designers* do sistema, o mercado permanecerá limitado.

São, assim, necessárias ferramentas que permitam aos potenciais utilizadores adaptar e criar um sistema inicial em dias ou semanas e não em meses.

A questão básica que se põe ao desenvolvimento de tal ferramenta é a forma e o nível de informação desejado pelo utilizador. Uma das possibilidades é produzir uma representação gráfica dos modelos, mas esta solução expõem muitos detalhes dos modelos. Em vez disso, muitos grupos estão a desenvolver sistemas que obtêm informação principalmente de exemplos de frases de interesse e de informação a ser extraída.

A.2 *Desempenho*

Uma outra barreira à proliferação do uso de sistemas de extracção é a limitação do desempenho. O que poderá contribuir para o aumento do desempenho? Em parte a convergência de tecnologias: os melhores sistemas existentes actualmente são praticamente semelhantes na sua apresentação global. Por outro lado estão as características do domínio. A experiência de outros fenómenos linguísticos parece indicar que uma grande fracção de dados relevantes estão codificados linguisticamente por um pequeno número de formas. É ainda de notar que o aumento da investigação nesta área provoca o aumento de cenários de extracção implementados. Assim, pode-se esperar ver conjuntos de modelos que são aplicados a famílias de cenários relacionados ou a domínios gerais. Por exemplo, modelos para acções básicas como compra e venda de produtos podem ser aplicados a muitos cenários dentro do domínio do negócio.

III. *Message Understanding Conferences*

Durante uma década[1] a IE foi conduzida por conferências para a compreensão de mensagens (MUC[2]). Estas conferências, instituidas pela *Defense Advanced Research Projects Agency* (DARPA) do ministério da Defesa dos Estados Unidos da América, ajudaram a formalizar a IE. Como exemplo, as tarefas de IE, denominadas *MUC tasks*, eram especificadas e incluíam a avaliação de critérios e *corpora* de texto para teste. As *MUC tasks* são ainda amplamente utilizadas para a avaliação dos sistemas de IE. Estas tarefas estão resumidas na tabela I.

Actualmente várias conferências da área da linguística computacional e da inteligência artificial lidam com a IE e com as suas sub-tarefas.

MUC	MUC *task*
MUC-1 (1987) e MUC-2 (1989)	mensagens sobre operações navais
MUC-3 (1991) e MUC-4 (1992)	artigos noticiosos sobre actividades terroristas
MUC-5 (1993)	artigos noticiosos sobre parcerias e microelectrónica
MUC-6 (1995)	artigos sobre mudanças de gerência
MUC-7 (1997)	noticias sobre veículose espaciais e lançamento de mísseis

TABELA I
RESUMO DAS TAREFAS APRESENTADAS NAS DIVERSAS MUC

Existem actualmente em investigação e desenvolvimento cinco tarefas de IE (IE *tasks*), definidas pelas MUC, que são:

- Reconhecimento de nomes de entidades (NE do inglês *Name Entity*) que encontra e classifica nomes, locais, etc.
- Resolução de co-referências (CO) que identifica relações de identidade entre entidades nos textos.
- Identificação de Elementos de *Template* (*Template Elements* (TE)) que adiciona informação descritiva aos resultados da NE (utilizando CO).
- Construção de Relações entre *Templates* (TR) que encontra relações entre entidades TE.
- Produção de *Templates* de Cenários (ST de *Scenario Template*) que enquadra os resultados da TE e TR em cenários de eventos específicos.

IV. SISTEMAS DE IE

A IE é realizada através de sistemas automatizados que extraem uma determinada informação pertinente de um grande volume de textos em linguagem natural. Extraem informaçao pré-definida sobre entidades e relações entre essas entidades, colocando-a num modelo de base de dados estruturados.

Os sistemas de IE têm sido desenvolvidos para um leque de estilos de escrita que vai desde o texto estruturado com a informação organizada de uma forma tabular até ao texto livre. O elemento chave para este último tipo de escrita é a definição de um conjunto de regras de extracção que identificam a informação relevante a ser extraída.

Para texto estruturado, as regras especificam uma ordem fixa de informação relevante e os *labels* delimitam os caracteres que devem ser extraídos. Para texto livre (o estilo de texto utilizado nos relatórios médicos), um sistema de IE necessita de várias outras ferramentas para além das regras de extracção. Nestas ferramentas estão incluídas as de análise sintáctica, as de *tagging* semântico, os reconhecedores de objectos do domínio, tais como pessoas e nomes de companhias, e as de processamento de discurso que fazem inferências para além dos limites das frases. As regras de extracção para texto livre são tipicamente baseadas em modelos que envolvem relações sintácticas entre palavras ou com as classes semânticas das palavras.

Uma outra característica importante de um sistema de IE diz respeito ao facto de este extrair apenas factos isolados de um texto ou ter a capacidade de relacionar informação e extrair múltiplos campos relacionados. Existem algumas áreas de análise em que a extração multi-campo é essencial. Existem, no entanto, outros domínios em que a extracção de campos singulares é perfeitamente adequada. No caso em que existe sempre menos de um evento por texto os campos podem ser identificados separadamente, e posteriormente tratados como um único caso.

Anteriormente foram apresentadas as MUC e as suas tarefas, as quais foram responsáveis pelo interesse inicial na IE. Foi principalmente a partir destas tarefas que surgiram os sistemas de IE mais utilizados. Alguns destes sistemas, bem como as suas principais características, estão sistematizados na tabela II.

Muitos outros sistemas de IE são construídos a partir de cascatas de autómatos de estados finitos. O sistema FASTUS (*Finite State Automa-based Text Understanding System*) da SRI *International*[3] é um destes sistemas. O FASTUS, financiado pelo DARPA, tem actualmente uma

[1] A primeira MUC (MUC-1) foi realizada em 1987, a ultima conferência MUC-7 foi realizada em 1997

[2] http://www.itl.nist.gov/iaui/894.02/related_projects/muc

[3] http://www.ai.sri.com/appeli/fastus.html

Nome	Estilo de Texto	Multi-campo	Sintaxe
WI	estruturado	sim	não
SRV	semi-estruturado	não	não
RAPIER	semi-estruturado	não	não
AutoSlog	livre	não	sim
CRYSTAL	livre	sim	sim
CRYSTAL	semi-estruturado	sim	sim
LIEP	livre	só	sim
WHISK	estruturado	sim	não
WHISK	semi-estruturado	sim	não
WHISK	livre	sim	sim

TABELA II
COMPARAÇÃO ENTRE SISTEMAS DE IE QUE UTILIZAM APRENDIZAGEM POR REGRAS

avaliação para o reconhecimento de nomes de 92% de *recall* e 96% de *precision* (próximo da *performance* humana). Para a extracção de informação a avaliação é de 44% e 61%, respectivamente. Algumas das principais características deste sistemas são a utilização de uma linguagem declarativa para a especificação de regras gramaticais e a aprendizagem automática a partir de modelos.

Existem outros projectos de IE em desenvolvimento actualmente. Destes destacam-se o ALEMBIC da Workbrench (MITRE)[4], o Highlight da SRI Cambridge[5], o LaSIE/Gate da Universidade de Sheffield (apresentado em mais detalhe na secção VII), o NetOwl da SRA [1], o IDENTIFINDER [2] (BBN) (reconhecedor de nomes baseado em HMMs), o PROTEUS da Universidade de Nova York[6] e, por exemplo, o TextPro de Doug Appelt[7].

V. IE NA MEDICINA

A IE pode ser útil numa grande variedade de domínios. As várias MUCs focaram domínios como, por exemplo, o terrorismo latino americano e a microelectrónica. No entanto, a extracção de informação médica é um asunto que também pertence ao domínio actual da IE.

Um dos primeiros sistemas a utilizar informação proveniente de relatórios médicos como domínio de aplicação foi o BADGER que utiliza o CRYSTAL [3] como algoritmo de extracção. O principal objectivo deste sistema passa pela análise de relatórios médicos e pela identificação de referências a "diagnósticos" e a "sinais ou sintomas".

Actualmente o domínio bio-médico é bastante utilizado para o desenvolvimento de sistemas de IE. Por exemplo, o *Medstract*[8], criado devido ao aumento significativo de nova informação biológica, permite aceder rapidamente a nova informação pertinente e obter, desta forma, uma ideia do último conhecimento biológico. O objectivo da *Medstract* é aplicar os avanços recentes em linguística computacional e em análise de textos na extracção de informação de grandes bases de dados de informação bio-médica como é o caso da MedLine.

O *Unified Medical Language System (UMLS)* [4], desenvolvido na National Library of Medicine (NLM)[9], facilita o desenvolvimento de sistemas computacionais que se comportem como se "percebessem" o significado da linguagem da biomedicina e da saúde. Com este propósito a NLM produziu e disponibiliza as UMLS *knowledge sources* (bases de dados) e ferramentas associadas (programas) para serem utilizados na criação de sistemas de informação electrónicos que criam, processam, recuperam, integram e/ou agregam dados e informação biomédica e de saúde. Estes não estão optimizados para nenhuma situação em particular, mas podem ser aplicados em sistemas que executam um conjunto de funções envolvendo um ou mais tipos de informação, por exemplo relatórios de pacientes, literatura cientifica directrizes e dados de saúde pública.

Outros exemplos de sistemas de IE para o domínio biomédico são os que utilizam o sistema de IE GATE. Por exemplo, desde Outubro de 2004 que a Medwrite Inc.[10] utiliza o GATE no seu software para aplicações médicas. Mas este não é o único projecto de domínio médico a utilizar o GATE. O *Enzyme and Metabolic Path Information Extraction* (EMPathIE)[11] foi um projecto desenvolvido pelos departamentos de Estudos de Informação e de Ciências de Computadores da Universidade de Sheffield, cujo objectivo passava pela aplicação das tecnologias de IE às tarefas bioinformáticas. O EMPathIE reutilizou muitas das componentes existentes no GATE e produziu outros módulos baseados nos utilizados em projectos relacionados, de modo a extrair detalhes de reacções de enzimas de jornais biomédicos.

VI. ARQUITECTURA DE UM SISTEMA DE IE

De uma forma geral, os sistemas de extracção de informação são constituídos por quatro módulos principais: um *tokenizer*, algum tipo de processamento lexical e morfológico, análise sintáctica e módulos específicos do domínio em análise que identificam a informação a encontrar numa aplicação particular.

No entanto, dependendo dos requisitos de uma aplicação em particular, é desejável adicionar módulos a este esqueleto.

O esquema de um sistema de IE generalizado é ilustrado no gráfico de fluxo da figura 1.

Nesta figura, cada rectângulo grande representa uma componente. As caixas cinzentas representam as componentes que não são utilizadas em todos os sistemas de IE e são opcionais. As setas ilustram o fluxo do trabalho de IE, as setas interrompidas representam os caminhos opcionais. Os rectângulos mais pequenos representam recursos tais como léxicos, *corpora* de texto, bases de dados, listas de expressões comuns e listas de palavras.

Primeiro, o *corpus* de texto é itemizado (*tokenised*) em parágrafos, frases e palavras. Após a itemização procuram-se todas as palavras num dicionário lexical e se necessário

[4]http://www.mitre.org/tech/alembic-workbench/workbench-overview.html
[5]http://www.cam.sri.com/html/highlight_demo.html
[6]http://nlp.cs.nyu.edu/
[7]http://www.ai.sri.com/%7Eappelt/TextPro/
[8]http://www.medstract.org/
[9]http://www.nlm.nih.gov/
[10]http://medwrite.biz/
[11]http://www.dcs.shef.ac.uk/research/groups/nlp/funded/empathie.html

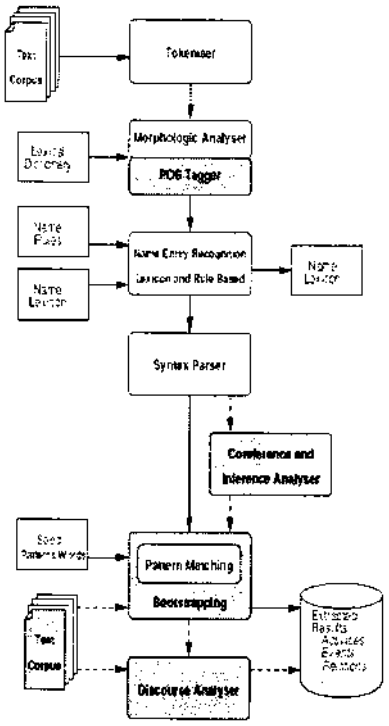

Fig. 1 - Esquema de um sistema de IE generalizado [5].

analisa-se a sua morfologia.

Alguns sistemas de IE aplicam na mesma fase os *Part-of-speech* (POS) *taggers*, com informação sintáctica adicional sobre as palavras, sob a forma de *tags*. A análise morfológica e o POS *tagging* estão amplamente relacionados e são frequentemente implementados como uma única componente.

A componente seguinte é denominada *Named Entity Recognition*. Os nomes das entidades consistem numa ou mais palavras e frequentemente representam informação a ser extraída. Existem vários métodos para o reconhecimento e extracção de nomes de entidades. Actualmente, o método mais aplicado é o reconhecimento de nomes de entidades com base em regras ou em léxico.

É possível adicionar mais informação sintáctica, comparável à informação proveniente da POS, através de *parsing* sintáctico. O *parsing* sintáctico não depende do reconhecimento de nomes de entidades e pode por isso ser aplicado também antes deste.

Os sistemas de IE recentes aplicam explicitamente análise de co-referência e inferências para produzirem melhores resultados. Alguns sistemas de IE pós-processam os resultados extraídos de modo a descobrir relacionamentos descritos no texto.

Encontrar modelos de extracção é a tarefa principal dos sistemas de IE. A informação é extraída utilizando estes modelos. Baseando-se na análise linguística efectuada sobre o texto, nas componentes descritas anteriormente, os modelos de extracção associam factos. Estes factos, ou peças de informação, são utilizados mais tarde para preencher os campos dos modelos dos resultados e são reunidos para formar tuplos de dados.

Os sistemas mais recentes usam vários algoritmos de *bootstrapping* para melhorar os resultados da associação de modelos, ou para fazer reconhecimento não supervisionado de nomes de entidades [5].

VII. GATE - *General Architecture for Text Engineering*

Esta secção descreve o sistema GATE, utilizado como software de base para o desenvolvimento do sistema de extracção de informação de relatórios médicos em Português.

O GATE é uma infraestrutura para o desenvolvimento e utilização de componentes de software que processam a linguagem humana. Em desenvolvimento na Universidade de Sheffield desde 1995, o GATE foi já utilizado numa grande variedade de pesquisas e projectos de investigação.

As componentes/recursos do GATE são tipos especializados de JavaBeans, mais especificamente, *Language Resources*, *Processing Resources* e *Visual Resources*.

Colectivamente, o conjunto de recursos integrados no GATE é denominado de CREOLE: *a Collection of REusable Objects for Language Engineering*.

Quando se utiliza o GATE para desenvolver funcionalidades de processamento de linguagem para uma aplicação, o investigador utiliza o ambiente de desenvolvimento para construir recursos dos três tipos. Isto pode envolver programação ou o desenvolvimento de Recursos Linguísticos tais como gramáticas que são utilizadas pelos "Processing Resources" existentes. É ainda possível uma mistura de ambas.

Quando um conjunto adequado de recursos tiverem sido desenvolvidos, podem ser embebidos na aplicação cliente alvo usando a estrutura do GATE.

A. *ANNIE* - A Nearly-New Information Extraction System

O GATE foi originalmente desenvolvido no contexto da investigação e desenvolvimento em IE. Vários sistemas de IE, em várias linguagens, de vários tamanhos e formas, foram criados utilizando o GATE com as componentes que foram distribuídas com este (ver [6] para a descrição de alguns destes projectos).

Uma família de *Processing Resources* para a análise linguística está incluída sob a forma de ANNIE, *A Nearly New Information Extraction System*.

Estas componentes utilizam técnicas de estados finitos para implementar várias tarefas desde a itemização até ao *tagging* semântico. Todas as componentes da ANNIE comunicam exclusivamente através dos documentos GATE e recursos de anotação.

O ANNIE é então uma família de recursos de processamento para análise linguística, tais como, *Tokeniser*, *Gazetteer*, *Sentence Splitter*, *Part-of-Speech Tagger*, *Semantic Tagger*, *Orthographic Coreference (OrthoMatcher)* e *Pronominal Coreference*.

B. *JAPE* - Java Annotation Patterns Engine

A linguagem JAPE permite o reconhecimento de expressões regulares sobre anotações em documentos.

Uma gramática JAPE consiste num conjunto de fases, em que cada uma é um conjunto de regras modelo/acção que

correm sequencialmente. Os modelos podem ser definidos pela descrição de um conjunto de caracteres específicos ou de anotações existentes (por exemplo, anotações criadas pelo *tokenizer, gazetteer, part-of-speech tagger*, ou pela análise do formato do documento). A definição de prioridades para as regras (se activada) previne a associação de multiplas anotações ao mesmo excerto de texto.

Até à data a JAPE foi utilizada com sucesso para o reconhecimento de nomes de entidades, *sentence splitting* e sumariação. Embora actualmente se utilizem regras produzidas manualmente, deve ser possível para uma aplicação aprender regras automaticamente.

As fases que constituem a gramática JAPE correm sequencialmente e constituem uma cascata de transdutores de estado finito sobre anotações. O lado esquerdo (LHS) das regras consiste em modelos de anotações que podem conter operadores de expressões regulares (ex.: *, ?, +). O lado direito (RHS) é constituído por expressões de manipulação de anotações. As anotações associadas pelo LHS de uma regra podem ser referidas no RHS através de etiquetas que são adicionadas aos elementos do modelo.

O exemplo seguinte apresenta uma regra gramatical para a extracção de um endereço de e-mail (assumindo uma definição apropriada de (EMAIL)) no caso de este ocorrer entre os sinais de menor e maior, respectivamente.

```
Rule: Emailaddress1
({Token.string == ''<''})
(
 (EMAIL)
)
:email
({Token.string == ''>''})
-->
 :email.Address= {kind = "email",
                  rule = "Emailaddress1"}
```

VIII. ADAPTAÇÃO AO PORTUGUÊS

Para o desenvolvimento do sistema de IE de relatórios médicos foi utilizado o GATE. Esta escolha é justificada pela sua simplicidade e capacidade de integração. No entanto, alguns dos recursos existentes neste sistema não são úteis para a extracção de informação em Português.

São as alterações já efectuadas a este sistema, com o objectivo de o adaptar à IE em Português que preenchem esta secção.

As primeiras experiências efectuadas com o GATE mostram que ferramentas como o *Tokeniser* e o *Sentence Splitter* podem ser utilizadas para a IE em Português. Assim, a tarefa de extracção de informação dos relatórios médicos começa com a utilização de um *Tokeniser* e de um *Sentence Splitter* cujos parâmetros/regras são as definidas no ANNIE.

Por outro lado, todos os resultados originados, quer pelo POS *tagger*, quer pelas listas de *Gazetteer* e mesmo pelas gramáticas semânticas, não correspondem, de uma forma geral, a resultados correctos. Este facto não é surpreendente uma vez que tais ferramentas baseam-se em conceitos linguísticos específicos da língua inglesa que não podem, por isso, ser aplicados ao Português.

A. *VMP Tagger*

Optou-se pela substituição do POS *Tagger* utilizado no ANNIE por um desenvolvido por Valentina Munóz em 2005 e disponível em *http://sourceforge.net/projects/vmptagger/*. Este POS *tagger* desenvolvido em Java e baseado no Brill Tagger [7], foi desenvolvido para a utilização integrada no GATE e para a categorização morfo-sintáctica em qualquer língua. Para tal apenas é necessária a especificação de quatro listas correspondentes a um léxico, um ficheiro de regras lexicais e outro de regras contextuais e um bigrama, que podem ser obtidas através do treino de pequenos *corpus* etiquetados na língua em análise.

As listas utilizadas neste trabalho são originárias de um etiquetador morfo-sintáctico desenvolvido na Universidade do Minho, que é descrito de seguida.

A.1 *Conjunto de* tags

Sendo a Língua Portuguesa de origem latina, tem uma relativa complexidade morfológica. A definição de um conjunto de *tags* para o Português é uma tarefa complicada e fulcral e corresponde também a um compromisso entre a precisão na descrição e a capacidade de aprendizagem de regras. Decidiu-se assim utilizar o conjunto de *tags* definido por Reis et al. [8]. Este conjunto segue alguns princípios funcionais tais como a precisão, a definição de uma estrutura hierárquica para cada etiqueta e baseia-se em modelos já utilizados em experiências com outras línguas.

A nomenclatura é de alguma forma hierárquica, o que implica que cada campo da *tag* tem um significado específico e estes campos são o mais reduzidos possível.

As *tags* são definidas com a seguinte estrutura:

Etiqueta: Categoria Sub-Categorias Género Pessoa Número

Categoria: QUE || SE || D || P || N || J || V || ADV || C || I || & || ?

referentes respectivamente a **que**, **se**, **D**eterminantes, **P**ronomes ou **P**reposições, **N**omes, ad**J**ectivos, **V**erbos, **ADV**érbios, **C**onjunções, **I**nterjeições, **C**ontracções e palavras desconhecidas.

As sub-categorias dependem, naturalmente, de cada categoria sendo os restantes elementos flexíveis dentro dos espaços conhecidos.

A categoria é o único campo obrigatoriamente preenchido. No caso de preposições ou interjeições será mesmo o único a ter valor.

B. *Gazetteer*

A alteração das listas do *Gazetteer* foi efectuada de modo a reflectir conceitos da língua portuguesa. Foram também adicionadas listas específicas do domínio, tais como listas correspondentes a nomes de doenças, a nomes de exames da área da Electroencefalografia, e a características dos resultados dos exames, entre outras.

C. *Gramáticas JAPE*

Para a extracção da informação do domínio pretendida foram desenvolvidas várias gramáticas baseadas na linguagem JAPE. Como exemplo referem-se as gramáticas para a

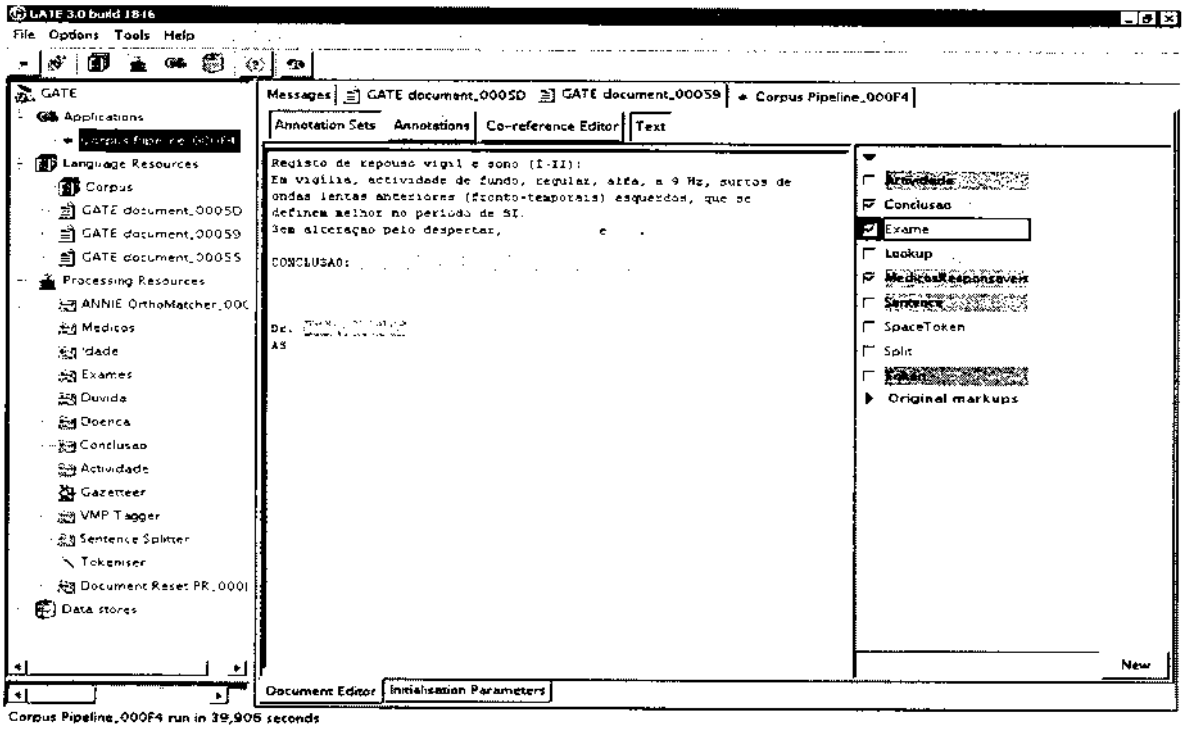

Fig. 2 - Elementos das entidades *MédicosResponsáveis*, *Conclusão* e *Exame*

extracção de nomes de doenças, características das actividades descritas nos relatórios, extracção de informação existente na conclusão destes, entre outras. Uma regra desenvolvida para a extracção do nome dos médicos responsáveis pelo exame, originando, deste modo, a entidade *Medicos-Responsaveis*, é apresentada de seguida.

```
/*
* When this grammar rule is invoked the
* rule Medicos is then invoked and this
* recognizes the lookup with majorType
* title (e.g. Dr.) followed by an space
* token. Following the rule Dr. Joao
* Lopes it will be annotated as a
* MedicosResponsaveis entity of type
* Name.
*/

Phase: medicos
Options: control = brill
Rule: Medicos
//e.g. Dr. Joao
({Lookup.majorType==title}
{SpaceToken}
)
(
{Token.orth == upperInitial}
({Token.string == "."})?
{SpaceToken}
{Token.orth == upperInitial}
)
:medicos -->
:medicos.MedicosResponsaveis
= {kind="name", rule="Medicos"}
```

IX. EXEMPLO

Um exemplo do resultado originado pela gramática anterior, em conjunto com todas as outras ferramentas apresentadas, pode ser analisado na figura 2.

Nesta figura é possível visualizar o ambiente de desenvolvimento do GATE bem como os recursos utilizados, em particular os *Language Resources* e os *Processing Resources* (lado esquerdo da figura). Na zona central encontra-se um exemplo de relatório utilizado e algumas das anotações efectuadas sobre este, concretamente as anotações que delimitam algumas das entidades consideradas importantes, como as entidades *MédicosResponsáveis*, *Conclusão* e *Exame*.

X. CONCLUSÃO

O sistema apresentado neste artigo resulta, quer de adaptações efectuadas a ferramentas de extracção de informação existentes no sistema GATE, quer da implementação/criação de novas ferramentas direccionadas para a IE em Português e de relatórios médicos da área da Neurofisiologia. São exemplo destas últimas algumas das gramáticas desenvolvidas.

Actualmente o sistema determina correctamente informação relativa ao conjunto de entidades consideradas relevantes no domínio. A inclusão de módulos que correlacionem as entidades identificadas, bem como o desenvolvimento de gramáticas para a resolução de anáforas, são algumas das tarefas a realizar brevemente, que poderão melhorar o desempenho do sistema.

BIBLIOGRAFIA

[1] Chinatsu Aone, Lauren Halverson, Tom Hampton, e Mila Ramos-Santacruz, "Sra: description of the ie2 system used for muc-7", em *Seventh Message Understanding Conference (MUC-7)*, San Francisco, California, 1998, Morgan Kaufmann Publishers.

[2] Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, Ralph Weischedel, e Annotation group, "Algorithms that learn to extract information; bbn: Description of the sift system as used for muc-7", em *Seventh Message Understanding Conference (MUC-7)*, San Francisco, California, 1998, Morgan Kaufmann Publishers.

[3] Stephen Soderland, David Fisher, Jonathan Aseltine, e Wendy Lehnert, "CRYSTAL: Inducing a conceptual dictionary", em *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Chris Mellish, Ed., San Francisco, 1995, pp. 1314–1319, Morgan Kaufmann.
URL: *citeseer.csail.mit.edu/soderland95crystal.html*

[4] D. Lindberg, B. Humphreys, e A. McCray, "Unified medical language systems", *Methods of Information in Medicine*, vol. 32, no. 4, pp. 281–291, 1993.

[5] Philipp Johannes Masche, "Multilingual information extraction", Master's thesis, Dept. of Computer Science, Faculty of Science, University of Helsinki, 2004.

[6] Diana Maynard, Hamish Cunningham, Kalina Bontcheva, Roberta Catizone, George Demetriou, Robert Gaizauskas, Oana Hamza, Mark Hepple, Patrick Herring, Brian Mitchell, Michael Oakes, Wim Peters, Andrea Setzer, Mark Stevenson, Valentin Tablan, Christian Ursu, e Yorick Wilks, "A survey of uses of gate", Relatório Técnico CS–00–06, Department of Computer Science, University of Sheffield, 2000.

[7] Eric Brill, "A simple rule-based part-of-speech tagger", em *Proceedings of ANLP-92, 3rd Conference on APlied Natural Language Processing*, Trento, Itália, 1992, pp. 152–155.

[8] Ricardo Reis e José João Dias de Almeida, "Etiquetador morfo-sintáctico para o português", em *Actas do XIII Encontro da Associação Portuguesa de Linguística*, Lisboa, Portugal, 1997, vol. 2, pp. 209–222, Associação Portuguesa de Linguística.