INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Clefomania 2

João Miguel Rodrigues da Cunha Guimarães

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática e de Computadores

Júri

Presidente:	Professora Doutora Ana Paiva
Orientador:	Professora Doutora Maria Luísa Torres Ribeiro Marques da Silva Coheur
Co-orientador:	Professor Doutor Nuno João Neves Mamede
Vogal:	Professor Doutor Bruno Martins

Julho 2009

Agradecimentos

À Professora Luísa Coheur e ao Professor Nuno Mamede pelo constante apoio e orientação durante a realização deste trabalho e pela disponibilidade demonstrada nas reuniões frequentes que tivemos.

Ao Professor David Matos pelas "luzes" que me foi dando ao longo do trabalho.

Aos meus pais e amigos pela força que me deram para que a realização deste trabalho fosse possível.

O meu muito obrigado.

Lisboa, 16 de Julho de 2009

João Miguel Rodrigues da Cunha Guimarães

Aos meus pais.

You know that children are
growing up when they start asking
questions that have answers.

Resumo

O QA@L^{2}F é um sistema de *question-answering* em desenvolvimento no Laboratório de Sistemas de Língua Falada, no INESC, desde Outubro de 2006 e caracteriza-se por ser formado por três componentes: pré-processamento do corpus; análise e interpretação da pergunta e extracção da resposta final. O objectivo do trabalho o desenvolvimento e a melhoria do sistema. Para esse efeito, foi realizado um estudo sobre as limitações do sistema na versão de Outubro de 2007 e implementado mais um módulo, a ser integrado com os restantes, de validação de resposta. Este módulo é composto por dois subcomponentes - o de validação do tipo da resposta e o de validação da fundamentação da resposta, tendo em conta um texto de suporte. Adicionalmente, implementou-se uma nova estratégia de extracção da resposta baseada em distâncias entre conceitos-chave presentes na pergunta e possíveis respostas candidatas. A técnica de Força Bruta sofre uma melhoria nomeadamente através da criação de novos padrões de detecção de dependências entre entidades mencionadas.

Abstract

QA@L^{2}F is a question-answering system developed at Laboratório de Sistemas de Língua Falada, at INESC, since October 2006 and is formed by three components: *corpora* pre-processing, question interpretation and answer extraction. The main goal of this project is to develop and improve the system by building, and integrate with the whole system, a new answer validation module. This module is made from two subcomponents, one that is responsible for answer type validation and the other that evaluates the veracity of the answer, according to a support text. Additionally, a new answer extraction technique was implemented. This new technique is based on distances between key concepts that are in the question and the candidate answers. Existing techniques like Brute Force technique are also improved, namely through the development of new dependencies patterns which represent relationships between named entities.

Palavras Chave
Keywords

Palavras Chave

Língua Natural

CLEF

AVE

Validação de Resposta

Padrões Linguísticos

Sistemas Pergunta-Resposta

Keywords

Natural Language

CLEF

AVE

Answer Validation

Linguistic Patterns

Question-Answering Systems

Índice

1 Introdução **1**

1.1 Motivação 1

1.2 Problemática 2

1.3 Objectivos do Trabalho 3

1.4 Estrutura da Tese 4

2 Problemas detectados em 2006 e Avaliação do sistema actual **5**

2.1 Introdução 5

2.2 Tipos de pergunta do CLEF 5

2.3 Descrição do sistema 6

2.3.1 Arquitectura 6

2.3.2 Estratégias na Extracção da Resposta 7

2.4 Versão em análise 9

2.4.1 Análise dos resultados 9

2.4.1.1 Perguntas do tipo *Factoid* 10

2.4.1.2 Perguntas do tipo *Definition* 12

2.4.1.3 Perguntas do tipo List 14

3 Estado da arte **15**

3.1 Introdução 15

3.2 Sistemas de QA em análise 15

3.2.1 Priberam 16

3.2.2 Joost 18

3.2.3 QUANTICO 19
3.3 Sistemas de Validação da Resposta 20
3.3.1 Sistema de Validação do INAOE 21
3.3.2 Sistema de Validação da UNED 23
3.3.3 Sistema de Validação da UAIC 25
3.3.4 Sistema de Validação do LIMSI 27
3.3.5 Outros sistemas de validação 29
3.3.6 Comparação entre sistemas de validação 30

4 Melhorias efectuadas no sistema **33**
4.1 Introdução 33
4.2 Melhorias na técnica de Força Bruta 33
4.2.1 Produção de dependências 33
4.2.2 Escolha da resposta por dependências 35
4.3 Técnica de extracção baseada em distâncias 36
4.3.1 Extracção de informação através de expressões regulares 40
4.4 Módulo de Validação da Resposta 41
4.4.1 Visão Global 41
4.4.2 Validação do tipo da resposta 42
4.4.3 Verificação da fundamentação da resposta 44
4.4.3.1 Cálculo do rácio entre conceitos patentes na pergunta e na resposta . . . 44
4.4.3.2 Cálculo das distâncias entre resposta candidata e entidades da pergunta 46
4.4.4 Cálculo da pontuação final 46

5 Avaliação **49**
5.1 Introdução 49
5.2 Avaliação do sistema de QA 51
5.3 Avaliação do módulo de validação da resposta 55

6 Conclusão **59**

I Apêndice 63

A Resultados da Avaliação Clef 2008 **65**

Lista de Figuras

1.1 Visão abstracta da arquitectura de um sistema QA 1

2.1 Arquitectura do QA@L^{2}F 7

4.1 Visão global do módulo de validação da resposta. 42

4.2 Classes e sub-classes de tipos de resposta 43

A Resultados da Avaliação Clef 2008 **65**

1 Introdução

1.1 Motivação

A grande motivação que se encontra por trás do desenvolvimento de sistemas de QA - Question Answering - reside na necessidade de se criarem mecanismos inteligentes capazes de seleccionar e extrair, de um conjunto grande de informação e num intervalo de tempo considerado útil, o conteúdo relevante que se pretende obter. O sistema QA típico responde a esta necessidade, possibilitando que um utilizador realize uma pergunta em língua natural. A Figura 1.1 esquematiza uma visão abstracta do

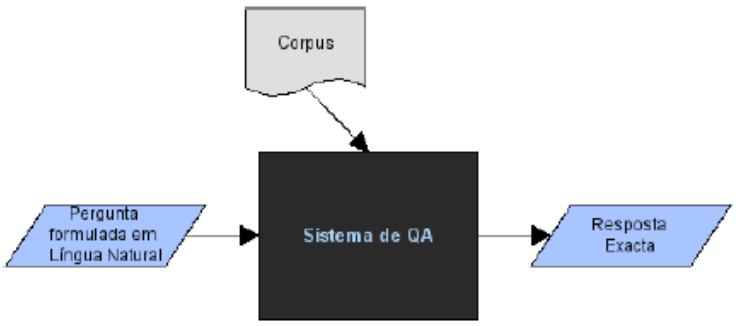

Figura 1.1: Visão abstracta da arquitectura de um sistema QA

funcionamento de um sistema de QA convencional. Cabe ao sistema de QA interpretar a pergunta, isto é, compreender que informação ela pretende obter e ir procurar tal informação na base de conhecimento acessível ao sistema de QA, tipicamente conjuntos de texto jornalísticos ou outras bases de conhecimento como a Wikipédia, e, finalmente, retornar essa informação exacta considerada pelo sistema como uma resposta à pergunta previamente efectuada.

Este trabalho tem como motivação dar continuidade àquele que foi realizado até Outubro de 2007, melhorando o comportamento do sistema QA@L^{2}F - o sistema de pergunta-resposta do L^{2}F - e dotando-

o de novas funcionalidades com o objectivo de apresentar melhores resultados no forum de avaliação CLEF na edição de 2008. As novas funcionalidades que se pretendem incorporar no referido sistema encontram-se referenciadas na secção 1.3 e explicadas, com maior detalhe, no capítulo 4.

1.2 Problemática

O grande desafio que se esconde por de trás da tarefa de QA é contornar as dificuldades impostas pela complexidade patente no Processamento de Língua Natural (doravante PLN). Mesmo para nós humanos, é complicado e exige certo esforço mental resolver ambiguidades, seleccionar a informação que nos interessa e interpretar e recolher informação implícita. Torna-se ainda mais problemático dotar máquinas e sistemas de QA com tal capacidade de raciocínio com o intuito de contornar as referidas dificuldades.

Por exemplo, na pergunta "*Como se chama o pai de João Loureiro?*" um ser humano consegue facilmente responder se tiver a informação de que "João Loureiro é filho de Valentim Loureiro..." através de um simples exercício de raciocínio em que se conclui que se X é filho de Y então Y é pai de X. No entanto a informação na frase de suporte não responde directamente à pergunta, pelo que lidar com informação implícita é um dos grandes desafios dos sistemas de QA.

Na pergunta "*O que é um brigadeiro?*" está-se perante um claro problema de ambiguidade. Que definição se pretende obter? Brigadeiro, doce de chocolate, ou brigadeiro patente militar? Mesmo para um ser humano tal ambiguidade pode ser de difícil resolução. Ou ele percebe, através do contexto no qual é feita a pergunta, a que brigadeiro ela se refere, ou restringe o domínio da resposta devolvendo uma pergunta do género "*Mas que tipo de brigadeiro?*".

Outra dificuldade relacionada com o processamento de Língua Natural reside no facto de se poder formular de diversas maneiras a mesma pergunta. Um sistema de QA deve recolher a mesma informação não obstante esta se poder apresentar sob diferentes formas. Por exemplo, as seguintes questões, formuladas de distintas maneiras, pretendem obter a mesma resposta:

```
Quem descobriu a América?
A América foi descoberta por quem?
Por quem foi descoberta a América?
```

Pode-se concluir que, por todas estas dificuldades, a tarefa de QA é complexa e exige um trabalho contínuo ao longo de vários anos. Tendo tal facto em consideração, este trabalho dá continuidade àquele que foi desenvolvido no primeiro ano do sistema QA@L^{2}F, visando a correcção de certos aspectos e a implementação de novas funcionalidades.

1.3 Objectivos do Trabalho

Tratando-se de uma continuação do trabalho até agora realizado, pretende-se delinear objectivos que visem a melhoria substancial do sistema QA@L^{2}F. Em primeira instância este trabalho requer uma análise profunda da versão do sistema de Outubro de 2007, visto ser a partir desta que este trabalho se inicia. A referida versão denota certas limitações e problemas:

- impossibilidade de responder a perguntas com anáforas,
- baixa taxa de acerto em perguntas do tipo *Factoid* - perguntas cujas respostas são entidades mencionadas de um determinado tipo. Por exemplo: Quem foi D. Afonso Henriques?
- respostas retornadas pelo sistema de um tipo diferente do esperado.

Não sendo um objectivo deste trabalho, pretende-se, todavia, aproveitar o trabalho a realizar por outros colegas de mestrado para se introduzir no QA@L^{2}F, tanto na fase de análise e interpretação da pergunta como nos textos jornalísticos, o tratamento de figuras de estilo como as anáforas e elipses de modo a que o espectro de perguntas que o sistema possa responder seja substancialmente alargado.

A baixa taxa de acerto para perguntas do tipo *Factoid*, pode ser parcialmente explicada pelo número reduzido de dependências detectadas entre entidades mencionadas. É com base nestas dependências que duas das técnicas de extracção da resposta, implementadas na versão do sistema de Outubro de 2007, escolhem possíveis respostas candidatas, pelo que se pode concluir que quantas mais dependências o sistema conseguir captar melhores resultados irá obter. Deste modo, torna-se fulcral melhorar a capacidade de o sistema produzir tais dependências através da melhoria e criação de novos padrões a serem implementados no XIP que se mostrem capazes de detectar relações entre pares de entidades mencionadas. Para tornar o sistema mais eficaz no tratamento de perguntas *Factoid*, o trabalho proposto vai adicionar ao sistema uma nova estratégia de extracção da resposta que se baseia no cálculo de distâncias entre vocábulos contidos na pergunta e entidades mencionadas no texto que são do tipo esperado na resposta, e que se tornam, por isso, respostas candidatas. Tal estratégia vai ser usada tanto para a Wikipédia como para os textos do *corpus* jornalístico disponível.

Outro problema verificado reside no facto das estratégias adoptadas no módulo de extracção da resposta final retornarem a resposta com base única e exclusivamente na ocorrência de vocábulos da pergunta que aparecem no texto de suporte. Este comportamento pode revelar-se bastante limitativo, dando motivação para que se implemente outros métodos como é o caso da atribuição de pesos às possíveis respostas que o sistema tem para escolher. Outro aspecto a melhorar, e aproveitando uma nova regra para o CLEF 2008, passa por retornar as 3 melhores perguntas candidatas, em vez de se retornar apenas a melhor.

Finalmente, é também proposto neste trabalho, sendo um seus focos principais, a implementação e respectiva integração, no sistema de QA, de um módulo de validação da resposta de modo a impossibilitar que sejam retornadas respostas de tipos diferentes ao esperado e a permitir uma classificação heurística ao grau de fundamentação da resposta, isto é, determinar a qualidade, com base em em métricas típicas usadas em sistemas de validação de resposta, do texto de suporte face à pergunta e respectiva resposta candidata.

1.4 Estrutura da Tese

O documento encontra-se estruturado da seguinte forma: no capítulo 2 faz-se uma breve descrição do funcionamento do sistema QA@L^{2}F, considerando a versão datada de Outubro de 2007, a partir da qual este trabalho se inicia, e são escrutinados os diversos problemas e limitações que tal versão apresenta. No capítulo 3 são descritos três sistemas de QA que têm vindo a apresentar resultados satisfatórios nas ultimas edições do fórum de avaliação CLEF e que contemplam técnicas consideradas interessantes e inspiradoras para o trabalho realizado no sistema QA@L^{2}F. Visto que um dos objectivos do projecto (1.3) passa pela implementação de um módulo de validação da resposta, este capítulo aborda também o trabalho relacionado com sistemas de validação de resposta, descrevendo aqueles que tiveram participação no AVE nos anos de 2007 e 2008. O capítulo 4 foca as melhorias efectuadas ao sistema entre Outubro de 2007 e Outubro de 2008. No capítulo 5 é feita uma avaliação ao sistema e, finalmente, no capítulo 6 são feitas a conclusões finais e recomenda-se o trabalho futuro a desenvolver.

2 Problemas detectados em 2006 e Avaliação do sistema actual

2.1 Introdução

Visto que este trabalho se trata de uma continuação de uma tese de mestrado do ano passado, onde foram descritos os mais importantes sistemas de *question-answering* desenvolvidos até à data, justifica-se uma abordagem que visa, em primeira instância, analisar o sistema QA@L2F (Mendes, 2007) desenvolvido nos anos de 2006 e 2007. Pretende-se determinar em que circunstâncias o sistema se desvia do comportamento esperado e perceber as razões que levam a tais situações e perceber qual o estado actual do sistema, escrutinando os seus mais importantes problemas e limitações, para que seja mais fácil delinear o trabalho futuro a realizar.

Como resultados obtidos no já referido fórum de avaliação, num total de 200 perguntas, o sistema conseguiu responder correctamente a 22, numa primeira submissão, e a 26 numa segunda. De referir que do total das 200 perguntas submetidas, o sistema só possuia o suporte necessário para tentar responder a 114. Esta baixa abrangência deve-se ao facto de certos aspectos como a presença de anáforas e restrições temporais não estarem ainda tratados no sistema QA@L2F. Outro factor que influência negativamente o comportamento do sistema, é o facto de apenas 30% do corpus se encontrar processado. No entanto, existem perguntas que, aparentemente, o sistema deveria responder correctamente e que, por variadíssimos motivos, tal não acontece. Tais perguntas, representativas dos vários tipos de pergunta que o sistema reconhece, passam agora a merecer uma análise profunda, com o objectivo de se perceber, especificamente, onde é que o sistema está a falhar.

2.2 Tipos de pergunta do CLEF

Quanto à classificação do tipo de perguntas usada na avaliação no CLEF 2007, têm-se em consideração quatro grupos:

- Factoid - Perguntas que se cingem a factos, isto é, perguntas do tipo “Quem”, “Onde”, “Quando”, “Quanto”.
  (ex.: `Quem foi D. Afonso Henriques?`);

- Definition - Perguntas que têm como resposta uma dada definição de um objecto ou de uma sigla. (ex.: `O que é a defesa siciliana?`);
- List - A resposta a este tipo de pergunta pode ser vista como uma lista em que cada elemento é uma sub resposta da pergunta.
  (ex.: `Diga o nome de três escritores portugueses.`);
- NIL - Perguntas cujo valor esperado na resposta é NIL.
  (ex.: `Que mar banha Braga?`).

Para cada tipo de perguntas o sistema tem ao ser dispor um conjunto de *scripts* de extracção de resposta que os chama consoante o resultado obtido na análise e interpretação da pergunta. É gerado um *script* que por sua vez tem indicado qual o *script* a chamar e os respectivos argumentos para a extracção da resposta.

2.3 Descrição do sistema

O QA@L^{2}F é um sistema de pergunta resposta actualmente a ser desenvolvido no Laboratório de Língua Falada. Tem como principal motivação conseguir responder correctamente a um cada vez maior leque de perguntas recebidas como input do sistema. Para estimular o desenvolvimento do sistema e facilitar a sua avaliação, revelou-se de extrema importância a participação no fórum CLEF - *Cross Language Evaluation Forum*.

2.3.1 Arquitectura

No seu primeiro ano de desenvolvimento, foi desenhada e implementada a arquitectura nuclear do sistema. Deste modo, consideram-se três importantes componentes presentes no QA@L^{2}F:

- Pré-Processamento do Corpus;
- Análise e Interpretação da Pergunta;
- Extracção da Resposta Final.

A Figura 2.1 esquematiza a arquitectura do sistema. O "Pré-Processamento do Corpus" tem como principais funcionalidades extrair, de um determinado conjunto de texto, informação relevante que mais tarde possa ser útil para obter respostas a determinadas perguntas, identificando entidades mencionadas que são armazenadas em base de dados de uma forma organizada, consoante o tipo a que elas pertencem. Por exemplo, as entidades mencionadas do tipo *PEOPLE* são guardadas na tabela como o

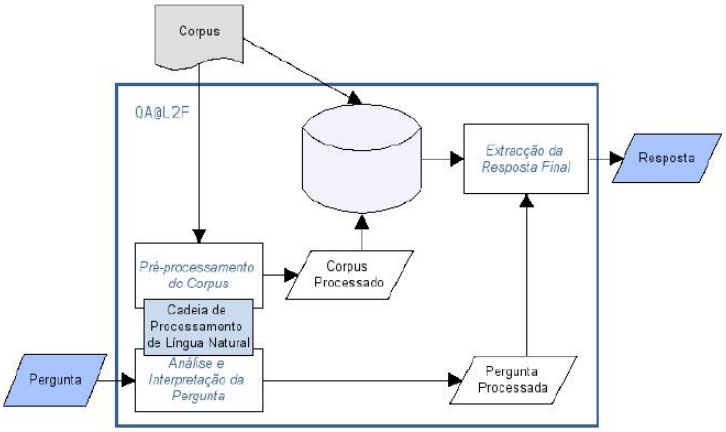

Figura 2.1: Arquitectura do QA@L^{2}F

nome *FACT_PEOPLE*. Na fase de "Análise e Interpretação da Pergunta", pretende-se recolher o máximo de informação presente na pergunta, tais como entidades mencionadas, tipo de pergunta e tipo de entidade mencionada que se espera na resposta. Ambos os componentes referidos utilizam a cadeia de PLN - Processamento de Língua Natural (Ribeiro et al., 2003; Pardal & Mamede, 2004; Ait-Mokhtar et al., 2001). Finalmente, depois da pergunta estar processada, é chamado o componente da "Extracção da Resposta Final" que, com base na informação obtida nos módulos anteriores, irá aplicar distintas estratégias para poder retornar uma resposta.

2.3.2 Estratégias na Extracção da Resposta

Nesta primeira versão encontravam-se implementadas no QA@L^{2}F quatro estratégias no módulo de "Extracção da Resposta Final" (Mendes, 2007):

- Emparelhamento de Padrões Linguísticos;
- Reordenação de Formulações Linguísticas;
- Emparelhamento de Entidades Mencionadas;
- Força Bruta com Pós-Processamento de Língua Natural.

O "Emparelhamento de Padrões Linguísticos" é uma técnica que consiste na interrogação directa a uma determinada tabela de facto do tipo idêntico ao da entidade mencionada captada na pergunta e considerada pelo o sistema como a mais importante.

```
Exemplo: Quem foi D. Afonso Henriques?
```

Neste exemplo, o sistema, na fase de análise e interpretação da pergunta, capta a entidade mencionada "D. Afonso Henriques", sabendo que se trata de uma entidade mencionada do tipo *PEOPLE*. Utilizando a técnica de extracção da resposta final em questão, o sistema interroga a tabela *FACT_PEOPLE* por "D. Afonso Henriques", tendo depois acesso a determinados campos capazes de retornarem a devida resposta.

Outra estratégia delineada é a "Reordenação de Formulações Linguísticas". A abordagem neste caso consiste em reformular uma determinada pergunta num ou mais padrões típicos de resposta.

```
Exemplo: O que significa X ?
Padrão típico: X é/foi/são/foram R.
```

Nas perguntas do tipo *Definition*, e não só, este método pode revelar-se bastante útil, reformulando a pergunta em padrões que, com grande probabilidade, podem aparecer no corpus.

Na estratégia de "Emparelhamento de Entidades Mencionadas", assume-se que a resposta está localizada no corpus perto de frases onde se verifica uma maior ocorrência de entidades mencionadas presentes na pergunta. São depois, nessas frases, procuradas e contabilizadas entidades mencionadas do tipo que se está à espera na resposta. A entidade mencionada que ocorreu com maior frequência é retornada pelo sistema como resposta final à pergunta. Se o QA@L^{2}F não consegue, através destas três estratégias, retornar uma resposta, utiliza uma outra estratégia designada "Força Bruta com Pós-Processamento de Língua Natural". Neste caso, o sistema interroga a base de dados do corpus não processado, utilizando informação recolhida na fase de análise e interpretação da pergunta. Como resultado da interrogação são devolvidas as frases do corpus onde há uma maior ocorrência de conceitos presentes na pergunta. Estas frases serão submetidas a um pós-processamento de língua natural, onde, em primeira instância, se tentam detectar dependências a partir de padrões implementados e embutidos na gramática do XIP. Por exemplo, considere-se o seguinte excerto de um texto:

```
"... em Paris (França)..."
```

O padrão é detectado pela seguinte regra:

```
| ?{?*, NOUN#1[location]}, PUNCT[paren, left],
NP{NOUN#2[location]}, PUNCT[paren, right] |
```

o que conduz à produção da dependência *LOCATION_OK(Paris, França)*, que nos dá a informação que Paris se situa na França. De salientar que é deste modo que as tabelas *FACT_* são preenchidas com

base nestas dependências aquando o pré-processamento do corpus. No entanto só 30% do corpus é que foi submetido a tal pré-processamento. Caso o sistema não conseguir responder através de uma dependência, procede-se à procura da entidade mencionada do tipo que se está à espera na resposta, que ocorre com maior frequência no restrito conjunto de frases consideradas. Se mesmo assim o sistema não for capaz de retornar uma reposta, então retorna o valor NIL.

2.4 Versão em análise

A versão do sistema, submetida a este estudo, não será aquela que foi usada no fórum de avaliação CLEF 2007. Tal política foi adoptada por duas razões:

- verificam-se constantes alterações na cadeia de PLN.
- no âmbito de elaboração de teses de mestrado, foram desenvolvidos novos sistemas e ferramentas - nomeadamente o Reconhecimento de Entidades Mencionadas (Loureiro e Romão, 2007) - cujas performances foram significativamente melhoradas depois da participação do QA@L2F no CLEF 2007.

Contudo, por uma questão de economia de tempo, o corpus pré processado no qual esta versão se baseia para efectuar a análise e interpretação da pergunta bem como a extracção da resposta, será o mesmo que o usado no referido fórum de avaliação. O corpus em questão contempla 30% dos artigos dos jornais Público e Folha de São Paulo nos anos de 1994 e 1995. Tendo como base as perguntas da primeira avaliação do CLEF 2007, o corpus pré processado e o sistema até então implementado, passa-se a estudar o comportamento do mesmo, analisando estatisticamente os sucessos e insucessos que as várias técnicas de extracção da resposta causam na performance do sistema. De referir que a versão sujeita a análise utiliza a cadeia de PLN de Outubro de 2007. Por essa razão a técnica de Força Bruta e o módulo de Análise e Interpretação da Resposta podem produzir resultados diferentes dos obtidos no CLEF 2007, realizado em Maio de 2007, visto que, ao efectuarem pós processamento em língua natural, utilizam a referida cadeia.

2.4.1 Análise dos resultados

São apresentados, nesta secção, os problemas mais comuns no tratamento de perguntas *Factoid* e de *Definition*. Embora pertencentes ao mesmo grupo de perguntas, o sistema revela comportamentos diferentes para os vários tipos de perguntas *Factoid*, pelo que se escolheu analisar separadamente três diferentes tipos - "Quem", "Quando" e "Onde". De realçar que perguntas do tipo "Quem é/foi/eram/são..."são consideradas como sendo do tipo Definition. As perguntas com anáforas foram descartadas desta avaliação, pois o sistema não tinha, até à data, um mecanismo de tratamento da referida figura de estilo.

2.4.1.1 Perguntas do tipo *Factoid*

Para perguntas do tipo "Quem" o sistema tenta, em primeiro lugar, uma abordagem baseada na técnica de Emparelhamento de Padrões Linguísticos (Mendes, 2007), onde se interroga a tabela de factos do tipo em questão e dá como resultado uma entidade do tipo *PEOPLE*.

```
P: Quem foi o 13º rei de Portugal?
R: Ângelo
Suporte: "...foi por o rei de Portugal enviado a Roma ,
onde longamente falou com Miguel Ângelo..."
Script: script-who-title.pl TARGET   "13º rei"
ENTIDADES "rei" TITLE "Portugal" LOCATION
```

Neste caso, o sistema falha, usando a técnica de Emparelhamento de Padrões Linguísticos, pois não encontra nenhuma entrada na interrogação à tabela de facto do tipo PEOPLE com o campo title "13º rei". Assim sendo, utilizando o Mecanismo de Relaxamento de Restrições (Mendes, 2007), o sistema vai procurar obter a resposta com base numa outra técnica - O Emparelhamento de Entidades Mencionadas. Deste modo, escolheu-se o documento de suporte à resposta que contém as entidades mencionadas presentes também na pergunta e a entidade mencionada do tipo de que se está à espera na resposta que, neste caso em particular, é do tipo *PEOPLE* (Miguel Ângelo). Como se pode verificar o sistema associou, erradamente, a entidade Miguel Ângelo às entidades do tipo *TITLE* (rei) e *LOCATION*(Portugal) presentes tanto na pergunta como no documento de suporte à resposta, o que indicia que esta técnica pode, em variadas situações, levar a comportamentos indesejados.

Para perguntas do tipo "Quando", o sistema tenta, numa primeira análise, obter a resposta, utilizando a técnica de Emparelhamento de Entidades Mencionadas de uma forma semelhante ao explicado no sub-capítulo anterior. Se não conseguir encontrar uma resposta utiliza a técnica denominada Força Bruta com pós processamento de Língua Natural, que consiste numa interrogação ao corpus não processado contendo os conceitos existentes na pergunta. As melhores frases devolvidas na referida interrogação são processadas e é extraída a entidade mencionada do tipo que se espera na resposta - neste caso *TIME* - que ocorre com maior frequência.

```
P: "Quando foi inaugurado o metro de Lisboa?"
R: Domingos
Suporte: Cactus: Cascais -- R. Domingos de Freitas,...
WHEN/script-when.pl TARGET VAZIO ENTIDADES "Lisboa " LOCATION
```

```
AUXILIARES "inaugurado"  "metro"
```

Nesta pergunta, mais uma vez, a resposta obtida não foi a esperada. O sistema tentou procurar no corpus processado, frases que contivessem a entidade mencionada "Lisboa" e as palavras auxiliares "inaugurado" e "metro". Como não obteve nenhum resultado redireccionou a tarefa para outro *script* responsável pela execução da Técnica de Força Bruta já descrita. Nesta fase, também se verificaram certos problemas visto que o sistema também não conseguiu encontrar no corpus não processado tais palavras-chave presentes na pergunta. Para resolver tal situação, o sistema é capaz de relaxar as interrogações à base de dados do corpus não processado, tornando-as mais abrangentes. Tal comportamento pode ser útil ao dar maior liberdade na decisão da resposta a extrair, no entanto, aumenta a probabilidade de o sistema retornar uma resposta incorrecta. É o que acontece neste exemplo, em que nova interrogação é feita às frases que contêm a entidade mencionada "Lisboa", desprezando-se as restantes palavras auxiliares. O documento escolhido como suporte à resposta é o que contém o maior número de referências à palavra "Lisboa", e é devolvida, como resposta, a palavra "Domingos" que pode ser categorizada como uma entidade mencionada do tipo *TIME* (Domingos - plural de Domingo).

O comportamento do sistema na extracção da resposta a perguntas do tipo "Onde" depende da interpretação feita na pergunta. Se a pergunta se incide sobre um local relacionado com um evento, é chamado um *script* que executa a técnica do Emparelhamento de Padrões Linguísticos e onde se procura no corpus processado as frases que contenham tal evento. Se a pergunta for mais simples, isto é, se se incidir apenas sobre a localização de um determinado local, é executada a técnica de Emparelhamento de Entidades Mencionadas, que vai interrogar directamente a tabela *FACT_LOCATION* para obter as possíveis respostas. Nos dois casos, se se não obtiver uma resposta é executada a técnica da força bruta. Tal acontece com a seguinte pergunta presente no fórum de avaliação:

```
P: "Onde fica o parque Eduardo VII?"
R: NIL
SCRIPT: script-where.pl TARGET  "parque Eduardo"
ENTIDADES "Eduardo " PEOPLE  "VII " NUM
```

Tratando-se de uma pergunta simples do tipo "Onde" em que se pretende saber uma localização mais genérica de "parque Eduardo VII", o sistema interroga a tabela *FACT_LOCATION* pelo campo *locationParent* do referido TARGET. No entanto, neste exemplo, verifica-se que tal interrogação não retorna nenhum resultado válido, pelo que é executada a técnica de força bruta com pós processamento de língua natural.

Num total de catorze perguntas *Factoid* do tipo "Quem" o sistema não conseguiu produzir nenhuma resposta correcta. Em seis perguntas (43%) foram detectados erros na fase de análise e interpretação da

pergunta. O sistema retornou repostas incorrectas para quatro perguntas (29%), todas elas na técnica de Força Bruta, e NIL para as restantes dez perguntas.

Para perguntas *Factoid* do tipo "Quando" o cenário é idêntico. Em nove perguntas o sistema responde correctamente a uma (11%), cuja resposta é NIL, e em quatro perguntas (44%) a fase de análise e interpretação da pergunta retornou informação incompleta. Em seis perguntas o sistema devolve NIL, resposta correcta para um dos casos, e em três (33%) devolve respostas incorrectas.

No forum de avaliação Clef 2007, só foram consideradas quatro perguntas do tipo "Onde", das quais uma o sistema respondeu correctamente. Para as restantes perguntas o sistema retornou NIL.

Constata-se assim que para perguntas da categoria *Factoid* o sistema revelou grandes limitações. Por um lado, a maior complexidade das perguntas, em comparação com perguntas do tipo *Definition*, levou a que se produzissem análises erradas na fase de interpretação da pergunta. Por outro lado, na fase de extracção de perguntas, as técnicas adoptadas não se revelaram eficazes na obtenção da resposta correcta. O trabalho futuro pretende melhorar significativamente os resultados nesta categoria, criando um maior e mais complexo número de padrões que permita a detecção de dependências na fase de processamento de língua natural e implementando novas alternativas de extracção da resposta utilizando, por exemplo, fontes de informação estruturada como a Wikipédia, só usada, no Clef 2007, para perguntas de categoria *Definition*. Um outro problema prende-se com o facto das politicas de relaxamento adoptadas originarem, na totalidade dos casos analisados, respostas incorrectas. Tal situação aconselha a que seja implementado e incorporado no sistema um mecanismo de validação da resposta final.

2.4.1.2 Perguntas do tipo *Definition*

Este é o tipo de perguntas onde o sistema apresentou melhores resultados, respondendo correctamente a 45% das perguntas na primeira submissão. A técnica utilizada para estes casos, Reformulação de Padrões Linguísticos, consiste em refazer a pergunta num padrão típico que se espera encontrar na resposta. A fonte de onde se extrai a resposta para perguntas de definição é a Wikipédia, que, devido à sua estrutura estandardizada, simplifica a tarefa do sistema em recolher a resposta final. Assim sendo, na análise e interpretação da pergunta é gerado um *script* que, passando-lhe como argumento o conceito do que se pretende obter a definição, invoca outro responsável por fazer a devida interrogação à base de dados da Wikipédia. A resposta obtida contempla o conjunto de caracteres presentes entre o padrão "`...Conceito...é/foi`" e o final da respectiva frase. Para perguntas de definição de abreviações, ou siglas, a abordagem seguida foi diferente. É utilizada a técnica de Emparelhamento de Entidades Mencionadas, interrogando directamente a tabela *FACT_STUFF*, onde estão guardados os acrónimos e os seus significados, encontrados aquando do processamento do corpus jornalístico. No entanto, tal

processamento pode levar a comportamentos indesejados. Atenta-se neste excerto de um corpus jornalístico:

```
Na quarta jornada, o Setúbal-FC Porto (RTP) e o
União de Leiria-Benfica (TVI)."
```

De acordo com as regras estabelecidas para detecção de abreviaturas, as siglas "RTP" e "TVI" são, errada e respectivamente, associadas a "Setúbal-FC Porto" e "União de Leiria-Benfica". A técnica de Reformulação de Padrões Linguísticos, usando a Wikipédia como suporte, também para perguntas de definição de siglas, pode ser uma alternativa viável em que certamente não se registarão tais ambiguidades. A seguinte tabela mostra os resultados obtidos para as 29 perguntas do tipo *Definition*, presentes na primeira avaliação do Clef 2007.

	N.º Perguntas	R	W	I	U
tipo Quem	10	3 (30%)	4 (40%)	3 (30%)	0 (0%)
tipo O que	19	13 (68,4%)	6 (31,6%)	0 (0%)	0 (0%)
Total	29	16 (55,2%)	10 (34,5%)	3 (10,3%)	0 (0%)

Tabela 2.1: Resultados CLEF 2007 - Perguntas Definition

De salientar que as perguntas do tipo Definition podem ser divididas em dois sub- tipos. As que visam um indivíduo e, que por isso começam por *"Quem é/foi/são...."*, e as que visam um objecto ou uma organização, que começam pelo padrão *"O que é/foi/são...."*. A grande diferença de comportamento na fase de extracção da resposta, reside no facto de que, para o primeiro sub-tipo, o sistema escolher em primeiro lugar a técnica de Emparelhamento de Padrões Linguísticos e só se tal técnica não retornar nenhum resultado é que irá utilizar a técnica de Reformulação de Padrões Linguísticos, utilizando a Wikipédia como suporte. Para perguntas de *Definition* do sub-tipo "O que", o sistema usa inicialmente a Reformulação de Padrões Linguísticos, abdicando da técnica de Emparelhamento de Padrões Linguísticos. Das dez perguntas do sub-tipo "Quem", o sistema responde correctamente a três, utilizando a técnica de Reformulação de Padrões Linguísticos com suporte dos textos da Wikipédia, e já depois de a técnica de Emparelhamento de Padrões Linguísticos não ter devolvido nenhuma resposta. Tal facto, aconselha a que, no futuro, mesmo para perguntas deste sub-tipo, seja evocada em primeiro lugar a técnica de Reformulação de Padrões Linguísticos. As quatro respostas erradas e as três inexactas demonstram comportamentos idênticos do sistema. Nestas sete perguntas o sistema não consegue retornar possíveis respostas nas tabelas de facto nem nas tabelas com informação da Wikipédia. É, por isso, chamada a técnica de Força Bruta que retorna resultados incorrectos. Das dezanove perguntas do

sub-tipo "O que" o sistema responde correctamente a dez, tendo como suporte os textos da Wikipédia, e a três, recorrendo à técnica de Emparelhamento de Padrões Linguísticos. De notar que estas três perguntas questionam a definição de abreviaturas, para as quais foram definidos padrões específicos para a sua detecção, aquando da realização do pré processamento do corpus. Na tabela `FACT\_STUFF` são guardadas as siglas bem como os seus respectivos significados por extenso. Das seis respostas erradas, o sistema retornou, usando a técnica de Força Bruta respostas incorrectas para quatro perguntas. Nas restantes duas respostas falhadas, observou-se uma falha na fase de análise e interpretação da pergunta, onde foram ignorados conceitos cruciais para adequadas interrogações à base de dados.

Analisando tais dados, chega-se à conclusão que:

- A técnica de de Emparelhamento de Padrões Linguísticos funciona mal para perguntas de definição não direccionadas a abreviaturas/siglas;
- Para perguntas que visam um indivíduo com mais de um nome, ou, genericamente, um conceito composto por mais de dois vocábulos, deve-se interrogar a base de dados por páginas da Wikipédia com e sem o caracter "_"entre os nomes que compõem o conceito em questão;
- A técnica da Força Bruta revela muitas debilidades, pois selecciona, como possíveis respostas, única e exclusivamente entidades mencionadas. Não possui nenhum mecanismo de análise sintáctica ou de detecção de padrões o que leva a que o sistema retorne respostas incompletas. As políticas de relaxamento, activadas quando o sistema não consegue retornar uma resposta, são por vezes demasiado abrangentes, levando o sistema a escolher artigos de suporte indevidos.

2.4.1.3 Perguntas do tipo List

O sistema desenvolvido até à data não trata convenientemente este tipo de perguntas. A ideia subjacente para a resolução deste tipo de perguntas, passa por interrogar a base de dados com os artigos da Wikipédia com o objectivo de se descobrir padrões do tipo "*... é/foi/são...Conceito*". Devido à forma bem organizada desta fonte de informação, a resposta será o título da página do artigo em que tal padrão foi encontrado. Na fase de análise e interpretação da pergunta, é também recolhida a informação sobre o número de elementos que se pretende na lista da resposta, que indicará o número pretendido de resultados distintos devolvidos pela interrogação à base de dados.

3 Estado da arte

3.1 Introdução

Este trabalho incide no estudo que foi feito no âmbito de sistemas de *question-answering* e pretende ser útil, não só, no desenvolvimento do trabalho futuro, focando certos aspectos que podem ser melhorados no sistema QA@L^{2}F, como também, no estudo, mais abrangente, de certos problemas que afectam os actuais sistemas de *question-answering*. Para além disso focam-se algumas abordagens utilizadas em sistemas de validação de resposta, que podem servir de base para o novo módulo de validação de resposta a integrar no sistema QA@L^{2}F.

Verificando os resultados obtidos no fórum de avaliação e correlacionando-os com as técnicas adoptadas para a extracção das respostas, chega-se à conclusão que o sistema QA@L^{2}F comporta-se melhor para perguntas do tipo *Definition*, utilizando a técnica de "Reordenação de Formulações Linguísticas". Tal técnica, aproveitando a forma bem organizada e estruturada dos artigos da Wikipédia e baseando-se na detecção de padrões linguísticos, justifica tais resultados satisfatórios que sugerem, como trabalho futuro, um aprofundamento da mesma, estendendo-a a perguntas do tipo *Factoid*. No âmbito desse trabalho a realizar, passa-se a analisar, neste artigo, aquilo que, relacionado com técnicas de extracção de padrões e captação de dependências, se encontra implementado em vários sistemas de pergunta-resposta que marcaram presença no CLEF 2007.

Deste modo, na secção 3.2 analisam-se três sistemas de QA que apresentaram os melhores resultados nas suas respectivas línguas e que usam técnicas de detecção e emparelhamento de padrões para a extracção da resposta. Visto que este trabalho visa a criação de um novo módulo de validação da resposta, inclui-se neste capítulo a análise de sistemas que participaram no AVE nos anos de 2007 e 2008. Deste modo, na secção 3.3 descreve-se o que foi feito até à data no que diz respeito a técnicas de validação da resposta.

3.2 Sistemas de QA em análise

A análise dos sistemas efectuada nesta secção, tem em consideração apenas os ambientes monolingue dos vários sistemas em análise, independentemente do facto de estes abrangerem, ou não, outros

domínios. Escolheram-se para análise sistemas que apresentam os melhores resultados nas respectivas línguas em que se focam e que se baseiam em técnicas que se pretendem vir a ser exploradas no trabalho a realizar no sistema QA@L^{2}F em 2008.

3.2.1 Priberam

A Priberam (Cassan et al., 2006, 2007), participante do CLEF desde 2005, aposta no desenvolvimento e detecção de padrões, presentes tanto na pergunta como nos textos do corpus. Existem três tipos de padrões:

- Padrões de pergunta - categorizam a pergunta em questão, determinando o tipo de pergunta e o tipo de resposta que se pretende obter;
- Padrões de resposta - responsáveis por captar e categorizar, nas frases de documentos de suporte, possíveis respostas;
- Padrões pergunta-resposta - usados na extracção de respostas candidatas a uma pergunta específica, previamente categorizada;

O sistema, ao receber uma pergunta como *input*, determina qual padrão de pergunta ela corresponde, atribuindo-lhe uma categorização específica. São activados os padrões pergunta-resposta e seleccionados os documentos, previamente processados, que contemplam idêntica categorização. Tenta-se encontrar nas frases destes documentos os padrões pergunta-resposta activos que, por sua vez, seleccionam respostas candidatas. A Priberam tem ao seu dispor uma panóplia de recursos que possibilitam a criação dos referidos padrões. Para além de vários recursos lexicais, contendo uma vasta informação semântica e ontológica para cada unidade lexical, destaca-se o *software* desenvolvido pela empresa - o SintaGest (Amaral et al., 2005). Com o auxílio desta ferramenta são produzidas e testadas regras contextuais para reconhecimento de entidades mencionadas e outras expressões, bem como padrões dos vários tipos acima descritos. De salientar que tais padrões criados são muito mais poderosos que simples padrões de cadeia de caracteres, abrangendo outros aspectos para além do emparelhamento de expressões regulares, e escritos a partir de um conjunto de termos reconhecidos pela ferramenta SintaGest.

```
Question(FUNCTION)
    : Word(quem) Distance (0,3) Root(ser) AnyCat(Nprop, ENT) = 15
```

Este padrão de pergunta abrange questões a que o sistema classifica como sendo do tipo *FUNCTION*, como, por exemplo, "`Quem foi Jorge Sampaio?`". O termo *Word* indica a palavra exacta que

se pretende detectar, que no caso é "Quem", *Root* indica a palavra a considerar que tem como lema "ser", podendo ser uma qualquer conjugação de tal verbo, e *Distance(0,3)* indica que entre elas pode haver entre zero a três palavras. Depois desta sequência de palavras é expectável encontrar um conceito que esteja categorizado como sendo um nome próprio ou uma entidade mencionada - `AnyCat(Nprop, ENT).`

Passa-se a produzir o padrão pergunta-resposta associado:

```
Answer
    : Pivot \& AnyCat(Nprop, ENT) Root(ser) Definition With Ergonym? = 20
```

O termo *Pivot* sugere que o conceito a ser captado, neste caso uma entidade mencionada ou um nome próprio, tem de estar presente na pergunta. Tal conceito seguido de uma qualquer conjugação do verbo "*ser*" antecede a resposta - *Definition* - na qual pode conter, embora não obrigatoriamente, uma palavra que designe uma profissão, cargo ou função - *Ergonym*.

Como já foi referido, numa fase de pré-processamento do corpus, é retirada a cada frase informação sobre possíveis respostas que possam ter para determinado tipo de perguntas. Tal tarefa é realizada recorrendo a padrões de resposta. O exemplo mostra um padrão de resposta que, actuando num conjunto de frases, selecciona aquelas que podem conter resposta à pergunta de exemplo "`Quem é Jorge Sampaio?`":

```
Answer(FUNCTION)
    : QuestIdent(FUNCTION\_N) = 10
    : Ergonym = 10
```

As frases captadas, categorizadas como pertencendo ao tipo *FUNCTION*, têm a particularidade de conter um nome que é visto como um identificador de perguntas do tipo *FUNCTION*. São também consideradas palavras que designam profissões, cargos e funções. As respectivas frases são classificadas pelo sistema como possíveis fontes para respostas a perguntas do tipo *FUNCTION*.

Como é perceptível nos vários exemplos aqui transcritos, para cada padrão é atribuído um valor heurístico, com o objectivo de dar prioridade àqueles padrões em que se acredita serem mais fiáveis. São feitos também ajustes a estes pesos, nomeadamente incrementando unidades aquando a captação de termos opcionais e decrementando quando dois conceitos chave estão "distantes"um do outro, isto é, se entre eles estão mais palavras que o mínimo exigido.

O facto de uma questão poder corresponder a mais do que um padrão de pergunta, levando a que esta seja classificada com mais de uma categoria. O excesso de categorias, pertencentes a uma só pergunta, representava uma das principais causa para erros de extracção de respostas candidatas.

Parte do trabalho realizado em 2007 visou, entre outros, a resolução deste problema. A ideia subjacente é a de recolher ainda mais informação na fase de análise da pergunta. Deste modo, e com a ajuda da tecnologia desenvolvida até à data para o FLiP[1], para além de se categorizar a pergunta, extrai-se também a sua estrutura sintáctica bem como a função sintáctica dos pivots, seus constituintes. A nova informação recolhida ajuda na escolha de uma categoria específica para a classificação da pergunta e, por consequência, uma extracção de respostas candidatas mais fiável. A versão do sistema monolingue para a língua portuguesa acertou em 50% das perguntas na edição de 2007 do CLEF e em 63,5% na edição de 2008 do mesmo fórum de avaliação.

3.2.2 Joost

O Joost (Bouma et al., 2007), desenvolvido na universidade de Groningen e com participação activa desde o CLEF 2005, é um sistema de QA direccionado para a língua holandesa. À semelhança do que se passa noutros sistemas, é composto pelos seguintes componentes:

- Análise da questão;
- Recolha de documentos;
- Extracção e selecção da resposta.

Para além destes, foi também desenvolvido o módulo Qatar, responsável pela extracção directa das respostas em modo *off-line*, sem ter de ser necessário recorrer ao componente de recolha dos documentos que possam conter respostas candidatas. Todo o sistema, desde a análise da questão até à fase de extracção da resposta, assenta-se numa analise sintáctica providenciada pelo sistema Alpino, que alberga um vasto conjunto de dependências para a língua holandesa. Deste modo, ao ser submetida uma pergunta no sistema, ela é processada pelo Alpino e são captados o tipo da pergunta bem como os conceitos considerados relevantes. Se o tipo da pergunta for coincidente com uma das tabelas disponíveis no Qatar, previamente preenchidas aquando o processamento do corpus, o sistema vai considerar para a fase de extracção e selecção da resposta o conjunto de frases que o Qatar retornou. Caso contrário, tem em conta, as frases retornadas pelo componente responsável pela recolha dos documentos com respostas candidatas.

Depois de uma breve descrição do sistema, passam-se a focar, com maior profundidade, as técnicas adoptadas na utilização da informação sintáctica recolhida pelo sistema Alpino. Numa primeira fase, na análise da questão, são definidos um ou mais padrões para cada classe de pergunta. Um padrão de dependência é descrito como `(Head/HIx, Rel, Dep/DIx)` onde *Head* é o lema do constituinte

[1]Ferramentas para a Língua Portuguesa.

principal da relação, *Dep* o dependente da mesma e *HIx* e *DIx* índices para que se possam distinguir possíveis ocorrências repetidas do mesmo token. O sistema classifica uma determinada pergunta como *[classe_da_pergunta][argumentos]**. A classe da pergunta é determinada consoante o padrão a que ela corresponde e os possíveis argumentos são as palavras-chave presentes na pergunta.

```
(wat/W, wh, is/I)            (is/I, su, hoofstad/H)
(hoofstad/H, mod, van/V)     (van/V, obj1, Country/C)
```

O exemplo dado, retrata um padrão para a classe de perguntas "capital" onde perguntas do tipo "`Wat is de hoofstad van Portugal?`" (Qual a capital de Portugal?) são captadas. O sistema classifica esta pergunta de exemplo como *capital(Portugal)*.

Para além desta abordagem, seguiu-se, a partir de 2006, uma outra que utiliza o sistema Lucene[2], capaz de indexar a colectânea de documentos sobre vários aspectos linguísticos, nomeadamente POS, entidades mencionadas e relações de dependências. Tal sistema disponibiliza ainda uma linguagem própria para a formulação de *query's* usadas a partir de perguntas previa e sintacticamente analisadas. Com tanta informação disponível pelos sistemas Lucene e Alpino, tornou-se indispensável a implementação de um algoritmo que seleccionasse e pesasse as palavras-chave, e possíveis relações entre elas, a considerar para cada pergunta.

Foram também desenvolvidos vários tipos de padrões de resposta para a resolução de perguntas de *Definition* capazes de captar apostos, modificadores nominais, disjunções e complementos, e modificadores predicativos. Este sistema apresentou, no CLEF 2007, 24,5% de respostas correctas no ambiente monolingue Holandês-Holandês.

3.2.3 QUANTICO

Tal como acontece tipicamente noutros sistemas de QA, o sistema alemão QUANTICO (Sacaleanu et al., 2007) caracteriza-se por ter um componente que trata a análise da questão, outro para a recuperação de passagens nos textos e ainda outros dois para extracção e selecção da resposta. O componente responsável pela análise e interpretação da pergunta gera, para cada questão, um resultado em formato XML onde se encontram descriminados os seguintes objectos:

- q-type - tipo da pergunta;
- a-type - tipo da entidade mencionada de que se está à espera na resposta;
- q-focus - conceito central da pergunta;

[2]http://lucene.apache.org/java/docs/.

- restrições - possíveis restrições adicionais, como por exemplo restrições temporais.

A informação sintáctica é recolhida através do analisador *SMES* e a informação semântica através de *"syntactic constraints"* e duas bases de conhecimento, onde se fazem corresponder entidades lexicais a respectivos a-types (ex.: `cidade->LOCATION`). Seguindo esta estratégia e aproveitando uma característica da língua alemã, é feito uma extensão desta correspondência a léxicos originados a partir de outros já catalogados. Por exemplo, sendo *Stadt*(cidade) uma *LOCATION*, *HauptStadt*(capital) e *GrossStadt* (metrópole) também o são.

Na fase de recuperação de passagens, os textos, previamente processados, já têm incluída informação útil para a extracção da resposta, nomeadamente entidades mencionadas e padrões linguísticos. É gerada uma *query* de extracção de passagens de textos com restrições quanto a palavras-chave presentes na pergunta e tipo de entidades mencionadas que se espera na resposta.

Na fase de extracção da resposta são delineadas duas estratégias diferentes consoante o *q-type*, atribuído pelo módulo de análise da questão. Para perguntas *Factoid* pretende-se obter um determinado tipo de entidade mencionada na resposta. Para perguntas *Definition* o processo de extracção passa pela utilização de um padrão de resposta do tipo "*[Conceito][verbo de definição][.+]*", cuja intenção é apanhar o texto que define o conceito patente na pergunta.

O sistema monolingue desenvolvido para a língua alemã respondeu correctamente a 30% das perguntas no CLEF 2007 e a 37% na edição de 2008 do mesmo fórum de avaliação.

3.3 Sistemas de Validação da Resposta

Uma das tarefas propostas nas várias edições do CLEF é o o AVE - Answer Validation Exercise. Sucintamente, esta tarefa tem como *input* três argumentos:

- Pergunta p,
- Respostas candidatas[3] r,
- Texto de suporte t.

Os sistemas participantes devem, a partir do referido *input*, produzir como *output*:

- VALIDATED, se a resposta r à pergunta p é provada pelo texto de suporte t,
- SELECTED, para a melhor resposta r do conjunto das respostas marcadas como VALIDATED,

[3] Até 2006 era só considerada uma resposta.

- REJECTED, se a resposta r à pergunta p não é provada pelo texto de suporte t.

Passa-se agora a analisar os sistemas que utilizam as mais variadas técnicas de validação e selecção da resposta. Estes sistemas caracterizam-se por seguirem uma abordagem típica que passa inicialmente por gerar, a partir da pergunta e da resposta, uma frase afirmativa denominada hipótese e de seguida compará-la com o texto que suporta tal resposta. Nesta fase de comparação é considerado um conjunto de características, nomeadamente:

- sobreposições de entidades mencionadas, isto é, detecção de entidades mencionadas contidas tanto na hipótese como no texto de suporte,
- sobreposições de n-gramas,
- tamanho da maior subsequência comum (LCS - do inglês, ***L**ongest **C**ommon **S**ubsequence*).

Os sistemas estudados baseiam-se em aprendizagem automática (Moriceau et al., 2008; Castillo, 2008; Garcia-Cumbreras et al., 2007; Téllez-Valero & Luis Villaseñor-Pineda, 2007) e em estratégias baseadas em informação sintáctica (Moriceau et al., 2008; Ferrández et al., 2007; Iftene & Balahur-Dobrescu, 2008; Glõckner, 2007). Escolheram-se, para uma análise mais profunda, aqueles sistemas que apresentaram melhores resultados e que utilizam estratégias possíveis de serem aproveitadas, tendo em conta as ferramentas disponíveis, para o sistema no qual este trabalho se incide.

3.3.1 Sistema de Validação do INAOE

O sistema de validação da resposta desenvolvido pelo INAOE[4] (Téllez-Valero & Luis Villaseñor-Pineda, 2007) tem como base as abordagens típicas seguidas pela maior parte deste tipo de sistemas referidas na secção anterior.

É com base em características detectadas, entre uma hipótese e um texto de suporte, que o sistema decide se essa hipótese é vinculada ou não por esse texto de suporte, validando a resposta em caso afirmativo.

No caso específico do sistema do INAOE, são produzidas duas hipóteses para cada par pergunta-resposta. A primeira resulta da substituição do sintagma nominal que contém a expressão interrogativa pela resposta a considerar. Deste modo, considerando a pergunta *"How many inhabitants are there in Longyearbyen?"* e a resposta *"180 millions of inhabitants"*, a primeira hipótese gerada é *"180 millions of inhabitants are there in Longyearbyen."*. Para a produção da segunda hipótese, o sistema detecta o verbo

[4]Instituto Nacional de Astrofísica, Óptica e Electrónica, no México.

principal da primeira e troca os sintagmas nominais - *"in Longyearbyen are there 180 millions of inhabitants."*. De seguida, é feita uma análise para determinar se as duas hipóteses geradas são vinculadas no texto de suporte. O sistema do INAOE realiza este análise em duas etapas:

- detecção de sobreposições de termos;
- cálculo de sequências de sobreposições de termos.

A primeira etapa consiste unicamente em contar as ocorrências de palavras contidas tanto no texto como nas hipóteses. Obviamente que, tendo em conta que as duas hipóteses são constituídas pelos mesmos vocábulos, basta ao sistema contar sobreposições de termos referentes ao texto de suporte e a uma das duas hipóteses geradas. O sistema guarda então a percentagem dos nomes, verbos, adjectivos, advérbios, datas e números contidos na hipótese e no texto.

A segunda etapa baseia-se no cálculo do LCS, a partir do par (texto de suporte, primeira hipótese) e (texto de suporte, segunda hipótese), onde se extrai a maior subsequência. É guardado um valor resultante da divisão do comprimento do LCS pelo comprimento da hipótese. A decisão de se considerar válida ou não uma resposta é suportada por máquinas de suporte vectorial (SVMs - do inglês, **S**upport **V**ector **M**achines (Cristianini & Shawe-Taylor, 2000)), de modo semelhante ao que é feito em (Castillo, 2008), onde são considerados por uma ordem de relevância predefinida os valores dos traços acima descritos, nos quais se destacam a percentagem dos nomes sobrepostos e o comprimento do LCS por serem os mais relevantes. Outra contribuição importante presente neste sistema é a introdução de dois novos traços que incidem em restrições de respostas impostas pelo tipo e forma da pergunta:

- Valor booleano que indica se uma restrição imposta pela classe da pergunta é satisfeita. Tem valor verdadeiro se a classe semântica da resposta extraída corresponde ao que se estava à espera e falso caso contrário.
- Valor booleano que indica se uma restrição a um tipo específico é satisfeita. Para isso, o sistema tem que determinar o *target* específico a considerar. O sistema foca-se no sintagma nominal que tem a parte interrogativa e escolhe o vocábulo mais importante. Na pergunta *"How many inhabitants are there in Longyearbyen?"*, o sintagma nominal a considerar seria *"How many inhabitants"*, pelo que *"inhabitants"* seria o target escolhido. Tem valor verdadeiro se o *target* aparece imediatamente antes ou depois da resposta e falso caso contrário.

Os resultados apresentados por este sistema no AVE 2007 foram bastante satisfatórios alcançando os 52,91% de F-measure. Os resultados obtidos mostram também, com uma taxa de sucesso de 75%, que o sistema é bastante preciso em seleccionar a resposta correcta para uma questão, quando essa resposta existe no conjunto de respostas candidatas.

3.3.2 Sistema de Validação da UNED

Com participação activa no AVE desde 2007, o sistema desenvolvido na UNED[5] (Rodrigo et al., 2007, 2008) caracteriza-se por apenas considerar entidades mencionadas para fundamentar se uma determinada resposta a uma pergunta é comprovada por um texto de suporte. O sistema desenvolvido esta dividido em 4 fases:

- reconhecer as entidades mencionadas da pergunta e texto de suporte;
- determinar relações de inclusão[6] entre as entidades mencionadas previamente reconhecidas;
- decidir, com base no resultado apurado na fase anterior, a validação para cada triplo (pergunta; resposta; texto de suporte);
- seleccionar de uma resposta do conjunto de respostas validadas.

No primeiro módulo, é usado um reconhecedor de entidades mencionadas para marcar todo o tipo de nomes próprios, expressões temporais e expressões numéricas. Estas duas últimas são normalizadas para facilitar a tarefa de detecção de relação de inclusão. Decidiu-se também não fazer qualquer distinção entre os tipos das entidades mencionadas detectadas, desprezando-os e considerando as entidades como sendo de um único tipo. A razão para a tomada desta decisão reside essencialmente no facto do reconhecedor de entidades mencionadas, em muitos casos, confundir expressões temporais e expressões numéricas. Se se tivesse em conta o tipo das entidades mencionadas, quer as detectadas no texto de suporte quer as detectadas na hipótese previamente formulada, as mesmas deveriam ser do mesmo tipo para que o sistema conseguisse determinar uma relação de inclusão entre elas. No exemplo seguinte, o sistema detecta uma entidade mencionada e atribui-lhe um tipo incorrecto:

```
Texto de suporte: Iraque invadiu o Kuwait em TIMEX(Agosto_de_1990)
Hipótese: Iraque invadiu o Kuwait em \textbf{NUMEX(1990)}
```

Ao se fazer o emparelhamento entre as entidades mencionadas presentes na hipótese e no texto de suporte, e considerando os diferentes tipos de entidades, o sistema chega à conclusão que o referido texto não acarreta a hipótese visto que a expressão *"Agosto de 1990"* é detectada como expressão temporal enquanto que a expressão *"1990"*, embora contida na primeira, é vista pelo mesmo reconhecedor como expressão numérica, pelo que não existe relação de inclusão entre as duas entidades. Ao se desprezar o tipo das entidades mencionadas, está-se, no entanto, a perder informação que seria imprescindível

[5] **U**niversidade, **N**acional de **E**ducación a **D**istancia, Madrid.

[6] Considera-se que a entidade mencionada EM_1 contém uma entidade mencionada EM_2 se a cadeia de caracteres de EM_1 contém a cadeia de caracteres de EM_2 (http://pt.wikipedia.org/wiki/Acarretamento).

para a validação de tipos em perguntas que restringem a resposta a um tipo específico. Por exemplo, em perguntas do tipo "*Em que cidade fica X*", o sistema pode validar uma resposta com base no módulo de decisão de validação da resposta, descrito mais à frente, sem ter em consideração que a resposta validada seja efectivamente uma cidade.

Devido ao facto de certas entidades mencionadas poderem estar escritas de maneiras diferentes, o sistema assume que existe um emparelhamento entre duas entidades mencionadas se, calculando a distância de Levenshtein (Levenshtein, 1966), elas se diferenciarem em menos do que 20%. Deste modo, para palavras distintas como *Yasir*, *Yasser* e *Yaser*, que expressam a mesma entidade mencionada, embora escritas de maneira diferente, o sistema consegue detectar uma possível relação de inclusão.

O módulo de decisão de validação da resposta é responsável por gerar uma hipótese para cada par pergunta-resposta. Visto que o sistema considera apenas entidades mencionadas, a hipótese gerada é constituída somente pelo conjunto de entidades mencionadas detectadas na pergunta e na resposta candidata. O seguinte exemplo mostra, a partir de uma determinada pergunta e respectiva resposta candidata, as entidades mencionadas que o sistema detectou e a hipótese que conseguiu produzir:

```
Pergunta: Que país o EM(Iraque) invadiu em EM(1990)?
Resposta: EM(Kuwait)

Hipótese: EM(Iraque), EM(Kuwait), EM(1990)
```

O modulo de validação verifica nesta fase se o texto de suporte inclui todas as entidades mencionadas contidas na hipótese. Se tal não se verificar o sistema retorna *REJECTED* para o triplo (pergunta; resposta; texto de suporte) em questão.

O grande problema resultante desta abordagem reside no facto do sistema estar totalmente dependente do reconhecedor de entidades mencionadas. Se, por algum motivo, este reconhecedor falhar a marcação de uma entidade mencionada, o comportamento do sistema desvia-se do desejado. Por exemplo, considere-se a seguinte pergunta e respectiva frase de suporte:

```
Pergunta: What is the name of the national airline in EM(Italy)?
Suporte: Italy's national airline EM(Alitalia)...
```

Mesmo tendo como resposta candidata *Alitalia*, o sistema vai rejeitar tal resposta, visto que a hipótese gerada continha as entidades mencionadas *Italy* e *Alitalia* mas o texto de suporte apenas continha *Alitalia*, pois o reconhecedor não detectou na frase de suporte a entidade mencionada *Italy*.

Para contornar este problema, implementou-se em 2007 uma abordagem alternativa, na qual o sistema rejeita a resposta candidata se nenhum *token* ou sequência de *tokens* presentes no texto de suporte

não contem nenhuma entidade mencionada contida na hipótese. Isto é, continua-se a considerar somente as entidades mencionadas na pergunta e resposta candidatas para a geração da hipótese. Todavia, são tidos em conta todos os *tokens* do texto de suporte para determinar se tal texto inclui as entidades mencionadas contidas na hipótese.

Esta abordagem alternativa melhorou a medida-F do sistema obtida no AVE 2007 de 0,30 para 0,33 para a língua inglesa. O sistema apresenta melhores resultados para o espanhol com uma medida-F de 0,47 devido ao facto de o reconhecedor de entidades mencionadas usado - FreeLing - estar mais desenvolvido para a língua castelhana.

3.3.3 Sistema de Validação da UAIC

O sistema desenvolvido na UAIC[7] (Iftene & Balahur-Dobrescu, 2008) segue uma abordagem similar aos analisados em cima, contudo, e ao contrário do sistema implementado na UNED, descrito em 3.3.2, considera uma panóplia de tipos de entidades mencionadas e procede a uma análise dos mesmos, comparando o tipo da entidade mencionada da resposta com o tipo esperado.

O componente nuclear do sistema é o *Textual Entailment System* - TE - que tem como objectivo mapear cada entidade mencionada da árvore de dependências da hipótese numa entidade da árvore de dependências associada ao texto de suporte. Para cada mapeamento é atribuída uma nota que o qualifica. Para calcular o valor global de uma hipótese é feito o somatório de todos os valores parciais obtidos nos referidos mapeamentos. Um dos problemas observados na participação do sistema no AVE 2007 (Iftene & Balahur-Dobrescu, 2007), era o facto de este atribuir 0 a hipóteses que contivessem pelo menos uma entidade mencionada que não fizesse parte do texto de suporte, independentemente de outras entidades emparelharem na perfeição no mesmo texto. Esta regra, demasiado rígida, tornava impossível que respostas contidas nestas hipóteses fossem seleccionadas. Uma alteração feita em 2008, passou por modificar esta regra, calculando o valor heurístico da hipótese e assinalando-a com a marca *NE Problem*, indicando ao sistema que existe pelo menos uma entidade mencionada na hipótese que não está contida no texto de suporte. Se todas as hipóteses consideradas a uma determinada pergunta tiverem *NE Problem*, é validada a que tem um maior valor.

O sistema global apresenta as seguintes fases:

- Criação de um padrão para uma determinada pergunta a partir do tipo da mesma. Por exemplo para a pergunta *"What is the occupation of Richard Clayderman?"* o padrão gerado é "The occupation of Richard Clayderman is JOB"em que JOB é uma variável;

[7]Universidade Alexandru Ioan Cuza, Roménia.

- Construção de um conjunto de hipóteses formado pelo padrão criado na fase anterior e as respostas candidatas;
- Execução do sistema TE descrito no parágrafo anterior para os pares (T, H_1), (T, H_2),..., (T, H_k), sendo T o texto de suporte e k o número de hipóteses formuladas na fase anterior.

Em 2008 o trabalho foi expandido, focando-se na validação dos tipos da resposta, onde foram adicionadas mais duas fases:

- Identificação do tipo da entidade mencionada das respostas,
- Identificação do tipo das entidades mencionadas esperadas para as perguntas.

O objectivo é de evitar que a resposta seleccionada seja de um tipo diferente do que se espera para a pergunta em questão. Foi usado o GATE[8] como reconhecedor de entidades mencionadas para profissões, cidades, países, localidades, pessoas e organizações e ainda desenvolvidos padrões para a detecção de datas e medidas. Se uma resposta não for reconhecida nem pelo o GATE nem por nenhum dos referidos padrões, ela é classificada como *OTHER*. Depois de devidamente classificadas as entidades mencionadas da resposta e de obtido o tipo esperado para a pergunta em questão, é somado, ao valor obtido no sistema TE, para cada hipótese, uma pontuação baseada no seguinte conjunto de regras:

- 1, se o tipo da pergunta e o tipo esperado para a resposta forem iguais;
- 1, se o tipo esperado para a pergunta for *"DEFINITION"* e o tipo da resposta *"OTHER"*;
- 0,5, se o tipo da resposta e o esperado para a pergunta fazem parte da mesma classe de tipos de entidades, a saber:
  - {CITY, COUNTRY, REGION, LOCATION};
  - {YEAR, DATE};
  - {COUNT, MEASURE, YEAR}.
- 0,25, se o tipo da resposta ou o tipo esperado para a pergunta é *OTHER*.
- 0, caso contrário.

Foram feitas duas submissões para o AVE 2008, das quais na primeira se desprezou a identificação de tipos na resposta e de tipos esperados, não se procedendo à subsequente validação dos mesmos. Verificou-se uma ligeira melhoria nos resultados da segunda submissão onde tal validação foi tida em

[8]http://www.gate.ac.uk.

conta. Deste modo, o sistema apresentou, para a primeira submissão, a medida-F de 0,17 e 0,22 para inglês e romeno respectivamente e, na segunda, 0,19 e 0,23. Apesar de não se tratarem de resultados excepcionais, decidiu-se analisar este sistema, visto que dá uma contribuição útil, nomeadamente na abordagem utilizada para a validação de tipos na resposta.

3.3.4 Sistema de Validação do LIMSI

O sistema do LIMSI[9] (Moriceau et al., 2008), desenvolvido em França no CNRS[10], consiste num sistema de QA vocacionado para a língua francesa que integra um módulo de validação da resposta. Na sua participação no AVE 2008 foram testadas e avaliadas duas abordagens distintas:

- a partir de uma estratégia baseada na recolha de informação sintáctica, através da qual o sistema decide se um texto de suporte para uma determinada resposta é uma reformulação da respectiva pergunta;
- a partir de uma estratégia de aprendizagem automática, onde o sistema decide se a resposta é válida com base em diversas métricas.

Um dos sistemas QA desenvolvido - FIDJI[11] - procede a uma análise sintáctica das perguntas e dos textos de suporte, captando relações de dependências. Considerando o triplo pergunta Q, resposta A_{ave} e texto de suporte T, o sistema avalia se todas as dependências captadas em Q estão em T. Para além disso, é feita uma execução ao sistema FIDJI com a pergunta Q que produz uma resposta A_{fidji}, e se se verificar que A_{fidji} = A_{ave}, o sistema valida a resposta A_{ave} e considera-a justificada pelo texto de suporte T. Por exemplo, na pergunta "*Qui est Lionel Mathis?*" (Quem é Lionel Mathis?) são captadas as dependências *attribut(ANSWER, Mathis)* e *NNPR(Mathis, Lionel)* em que *ANSWER* é uma variável onde a resposta candidata se irá encaixar. No texto de suporte *"Lionel Mathis est un footballeur français, né le 4 octovre 1981..."*("Lionel Mathis é um jogador de futebol francês nascido em 4 de Outubro de 1981...") são produzidas, para além de outras, as dependências *attribut(footballeur, Mathis)* e *NPPR(Mathis, Lionel)*. O sistema FIDJI associa o lema "*footballeur*" à variável *ANSWER* e extrai, como resposta candidata, o sintagma nominal do referido vocábulo - "*footballeu français*". Visto que todas as dependências extraídas na pergunta são também extraídas no texto de suporte, a resposta submetida a validação é aceite se for igual à extraída pelo sistema FIDJI, ou seja, "*footballeu français*". O sistema FIDJI também é responsável pela marcação de entidades mencionadas, considerando 20 tipos diferentes, que, combinado com a análise da pergunta, permite obter o tipo da resposta retornada e o tipo esperado e verificar se são compatíveis.

[9] **L**aboratory for **M**echanics and **E**ngineering **S**ciences.
[10] **C**entre **N**ational de la **R**echerche **S**cientifique.
[11] **F**inding **I**n **D**ocuments **J**ustifications.

Posto isto, o sistema de validação de resposta, desenvolvido a partir do sistema FIDJI, valida uma resposta candidata se se verificar que:

- a resposta é também sugerida pelo próprio sistema FIDJI;
- o tipo da entidade mencionada da resposta é o esperado;
- a taxa de dependências não captadas tanto na pergunta como no texto de suporte é inferior a um limite específico previamente estipulado - 30%.

Como já foi referido, a segunda abordagem para a validação da resposta baseia-se num sistema QA de aprendizagem automática - FASQUES, onde são calculadas e consideradas várias métricas para a extracção de respostas. A primeira dessas métricas, vista como a mais importante, é a taxa de termos da pergunta contidos também no texto de suporte. São tidos em consideração quatro classes de termos:

- *Focus* (idêntico ao *target* no QA@L^{2}F), a entidade na qual a pergunta se incide. Por exemplo, na pergunta *"De que partido político é Lionel Jouspin?"* tem como focus o termo "*Lionel Jouspin*";
- tipo da resposta (idêntico ao *target-type* no QA@L^{2}F), se o tipo da resposta é explicitamente especificado. Na pergunta de exemplo anterior o tipo da pergunta seria *"partido político"*;
- o verbo principal da pergunta que possa corresponder a um facto ou acção;
- termos compostos por duas palavras sintácticamente ligadas como, por exemplo, *"Prémio Nobel"*.

Tal como na abordagem utilizada pelo sistema FIDJI, é considerado fundamental para a verificação da resposta a semelhança entre a resposta dada pelo sistema FRASQUES e a resposta submetida para validação. Para perguntas *Factoid* o FRASQUES selecciona a entidade mencionada do tipo esperado mais próxima dos termos considerados na questão. Outra métrica usada nesta abordagem passa pelo cálculo da maior sequência de conjunto de palavras consecutivas presentes no texto de suporte e na hipótese, construída através da concatenação da pergunta na sua forma afirmativa e da resposta candidata. A métrica a considerar resulta do rácio entre o número de palavras na cadeia obtida e o número de palavras da hipótese. Para facilitar a comparação entre os vocábulos do texto de suporte e da hipótese, os mesmos foram normalizados, considerando-se os seus lemas e sinónimos. O sistema FRASQUES aproveita também os seguintes traços originariamente calculados no sistema FIDJI:

- o resultado da validação feita pelo FIDJI, isto é, se a resposta foi validada, ignorada ou rejeitada pelo FIDJI;
- verificação se o tipo da entidade mencionada da resposta é compatível ou não com o esperado;
- taxa de dependências presentes na questão e em falta no texto de suporte.

Este sistema apresentou resultados bastantes satisfatórios no AVE 2008, mostrando ser aquele com melhores resultados para a língua francesa. Utilizando apenas o sistema FIDJI o sistema obteve uma medida-F de 0,57. Seguindo a abordagem do sistema FRASQUES os resultados melhoram, obtendo-se uma medida-F de 0,63.

3.3.5 Outros sistemas de validação

Os sistemas descritos na secção anterior seguem variadas abordagens típicas em sistemas de validação de resposta. Como já foi referido, conceptualmente existem dois grupos distintos de estratégias. Num encontram-se as estratégias que se caracterizam por se basearem em informação sintáctica recolhida a partir de, por exemplo, dependências ou inferências captadas e, o outro, contempla as estratégias de aprendizagem automática a partir de um conjunto de treino e de uma panóplia de métricas e traços que contribuem para a decisão da validação de uma terminada resposta. Alguns sistemas, como é o caso do sistema desenvolvido no LIMSI e descrito em 3.3.4, apresentam como solução dois sub-sistemas integrados que contemplam os dois tipos de abordagem. Para além dos sistemas analisados na secção anterior participaram no AVE2008 outros que merecem breves considerações e que se descrevem brevemente de seguida:

O sistema ProdicosAV (Jacquin et al., 2008), que não é mais de que um módulo do sistema de QA para a língua francesa Prodicos (Monceaux et al., 2005). A validação da resposta baseia-se em dois passos:

- validação temporal,
- validação da resposta, comparando a resposta submetida a validação com a resposta que o próprio sistema Prodicos retorna.

O sistema FaMAF (Castillo, 2008) segue uma abordagem de aprendizagem automática usando uma máquina de suporte vectorial, em que são tidas em conta 12 características para determinar a semelhança lexical entre o texto de suporte e a hipótese previamente formulada. Dos 12 traços destacam-se:

- percentagem das palavras da hipótese contidas no texto e vice-versa;
- percentagem de bi-gramas da hipótese no texto e vice-versa;
- distância de Levenshtein entre texto de hipótese;
- percentagem de trigramas da hipótese no texto e vice-versa;

- cálculo do TF-IDF[12], onde se procura obter a importância de um determinado vocábulo ou conjunto de vocábulos da hipótese no respectivo texto de suporte, e a similaridade do coseno[13] para determinar a semelhança entre hipótese e texto com base nos TF-IDF's previamente obtidos.

Numa segunda submissão para o AVE de 2008 foi ainda considerada o cálculo da semelhança semântica usando a Wordnet[14].

O sistema SINAI (Garcia-Cumbreras et al., 2007), que usa também abordagens típicas de aprendizagem automática, considerando um conjunto de métricas para calcular semelhança lexical entre texto e hipótese:

- emparelhamento simples onde é calculado a distancia semântica entre *tokens* da hipótese e texto. De entre várias medidas possíveis para o cálculo de tal distância, optou-se pela medida de similaridade de Lin (Lin, 1998) por apresentar melhores resultados;
- emparelhamento de subsequências entre hipótese e texto - CSS[15];
- emparelhamento de trigramas.

O sistema desenvolvido na universidade de Alicante (Ferrández et al., 2007), compara o texto de suporte com a hipótese criada a partir da pergunta e resposta candidata. Apresenta dois módulos distintos:

- módulo lexical que se baseia num leque de métricas semelhantes às usadas no sistema SINAI, tais como emparelhamento simples de vocábulos, distância de Levenshtein, emparelhamento de subsequências de vocábulos entre hipótese e texto - CSS - e emparelhamento de trigramas.
- módulo sintáctico que é responsável por gerar as árvores sintácticas da hipótese e texto de suporte, descartar informação considerada irrelevante nas duas árvores e fazer o emparelhamento nó a nó das duas árvores.

3.3.6 Comparação entre sistemas de validação

As tabelas seguintes indicam, sucintamente, para cada um dos 8 sistemas referenciados, que estratégias estes utilizam. Na Tabela 3.1 consideram-se técnicas comuns a sistemas baseados em aprendizagem automática tais como emparelhamento simples de vocábulos da hipótese e texto de suporte, a utilização do algoritmo LCS e do CSS, cálculo de distâncias Levenshtein entre vocábulos e geração e consequente emparelhamento de trigramas.

[12] *term frequency–inverse document frequency*.
[13] http://en.wikipedia.org/wiki/Cosine_similarity.
[14] http://wordnet.princeton.edu/.
[15] ***C**onsecutive **S**ubsequence*.

Sistema	Emp. Simples	LCS	CCS	Levenshtein	Trigramas
Alicante	✓		✓	✓	✓
FaMF	✓			✓	✓
INOE	✓	✓			
LIMSI	✓	✓			
ProdicoAV					
SINAI	✓		✓		✓
UAIC	✓				
UNED	✓			✓	

Tabela 3.1: Sistemas de Validação e técnicas baseadas em aprendizagem automática

Sistema	Dependências	Emp. de Árvores Sintácticas	Val. temporal	Comparação AVE e QA	Val. do Tipo
Alicante		✓			
FaMF					
INOE					✓
LIMSI	✓			✓	✓
ProdicoAV			✓	✓	
SINAI					
UAIC					✓
UNED					

Tabela 3.2: Sistemas de Validação e técnicas baseadas em informação sintáctica

Por sua vez, na tabela 3.2 explicitam-se que técnicas, baseadas na recolha de informação sintáctica e semântica, são usadas pelos diversos sistemas. As técnicas em causa são a construção e emparelhamento de dependências entre hipótese e texto de suporte; geração de árvores sintácticas e consequente emparelhamento entre hipótese e texto de suporte; validação de expressões temporais; comparação da resposta submetida a avaliação com a resposta dada pelo sistema QA e a validação do tipo de resposta, isto é, se o tipo da resposta candidata é compatível com o esperado.

4 Melhorias efectuadas no sistema

4.1 Introdução

No capítulo 2 são descritos e analisados vários tipos de problemas da versão do sistema QA@L^{2}F em Outubro de 2007. Este capítulo aborda detalhadamente as alterações e melhorias implementadas no ano lectivo de 2007/2008 que tentam ser solução para os referidos problemas, dotando o sistema de mais estratégias e melhorando outras já existentes, com o objectivo final de alargar o espectro de respostas correctas. Na secção 4.2 descrevem-se as melhorias feitas na técnica de Força Bruta, nomeadamente a produção de novas dependências sintácticas e introdução de pesos para a pontuar respostas candidatas. Na secção 4.3 explica-se uma nova técnica de extracção de resposta baseada em distâncias e, finalmente, na secção 4.4 é focado o módulo de validação da resposta.

4.2 Melhorias na técnica de Força Bruta

Um dos objectivos que este trabalho contemplou focou-se na melhoria da técnica de extracção de resposta denominada Força Bruta. De entre vários aspectos que levaram a tal melhoria destacam-se os seguintes que merecerão uma análise mais cuidada:

- A implementação de novos padrões linguísticos,
- A atribuição de pesos para respostas candidatas,
- O retorno de três respostas possíveis, aproveitando as novas normas do CLEF 2008.

4.2.1 Produção de dependências

Em 2.3.2, explica-se sucintamente a técnica de Força Bruta. Nesta secção aprofunda-se mais detalhadamente a técnica em questão e explicam-se os melhoramentos realizados. A técnica de Força Bruta tenta responder às perguntas através de duas alternativas: por entidades mencionadas do tipo que se espera na resposta, retornando aquela que ocorre mais vezes, ou por dependências detectadas por padrões linguísticos previamente definidos. O trabalho desenvolvido para o CLEF 2007 construiu importantes

alicerces para a exploração desta última abordagem, em que foram definidos alguns padrões mais simples e mais frequentes na língua portuguesa. Fez parte deste projecto dar continuidade ao trabalho desenvolvido neste tema específico, dotando o sistema de mais padrões linguísticos definidos e embutidos na ferramenta XIP, contida na cadeia de PLN. Deste modo, foram implementados 27 novos padrões linguísticos direccionados para perguntas do tipo "Quem", "Onde" e "Quando" e introduziram-se padrões mais complexos, que relacionam mais do que duas entidades mencionadas. Apresenta-se de seguida a regra definida no XIP para o padrões do tipo `<people>, nascido em <data>` que produz a dependência `DATE(<people>, <date>, nascido)`:

```
| ?{(PREP),(ART), NOUN#1[people]}, PUNCT[comma],
AP#3{PASTPART[lemma:nascer];PASTPART[lemma:matar];
PASTPART[lemma:falecer];PASTPART[lemma:assassinar]},
PP{PREP[lemma:a];PREP[lemma:em],NOUN#2[date]} |
DATE[OK=+](#1, #2, #3)
```

Devido ao facto do XIP em certos casos não juntar toda uma entidade mencionada num nó da sua árvore sintáctica, aproveitou-se esta possibilidade de adicionar mais argumentos a uma dependência para poder conter tal entidade, não perdendo assim informação. Contudo, esta opção obriga a que o sistema distingua os argumentos de uma dependência que sejam entidades mencionadas completas daqueles que representam parte de uma entidade mencionada fragmentada e que necessitam de ser copulados à restante expressão da respectiva entidade. Por exemplo, a entidade mencionada "presidente da Câmara de Lisboa" é dividida num NP - presidente - e num PP- "da Câmara de Lisboa". Posto isto, na pergunta "`Quem foi Jorge Sampaio?`", submetida directamente a esta técnica de extracção de resposta, o sistema detecta o padrão `PEOPLE_OK(Jorge Sampaio,presidente,de a Câmara de Lisboa)` com base no seguinte excerto de um artigo do corpus jornalístico:

```
...que será assinado pelo presidente da
Câmara de Lisboa, Jorge Sampaio,...
```

O sistema percebe que o 3º argumento desta dependência é a continuação do 2º, visto que começa com a preposição "de". É feita a concatenação dos dois argumentos e o sistema passa a tratar tal dependência como se tivesse apenas dois argumentos, relacionando "Jorge Sampaio" com "presidente da Câmara de Lisboa".

Outro problema verificado no tratamento dos resultados obtidos pelo XIP, diz respeito ao facto de, como seria expectável, nem todas as entidades mencionadas estarem devidamente marcadas. Tal facto

levou à criação de réplicas dos padrões linguísticos implementados mas desta feita com regras mais relaxadas. Isto explica o porquê do sistema contemplar, entre outras, dependências do tipo `LOCATION_OK` e `LOCATION_KO`. Como os próprios nomes indicam, na segunda existe muito maior incerteza quanto à veracidade da relação entre os seus elementos. O sistema toma tal situação em consideração e premeia as dependências do tipo *OK* no sistema de atribuição de pesos que passa a merecer uma análise mais profunda.

4.2.2 Escolha da resposta por dependências

Como já foi referido, foi desenvolvido um sistema de atribuição de pesos a respostas candidatas extraídas a partir da técnica de Força Bruta com base em dependências. Inicialmente, todas as respostas candidatas recolhidas têm um peso idêntico de uma unidade. Posteriormente são adicionadas unidades aos pesos de cada resposta candidata consoante se verifique a ocorrência de determinados factores considerados positivos. Com base nesta estratégia, quando se verifica que uma cadeia de caracteres que representa uma determinada entidade, considerada como *target* de uma pergunta, está totalmente contida no argumento de uma dependência, ou vice-versa, o peso da respectiva resposta candidata, que será o segundo argumento da dependência em questão, acresce em uma unidade. Deste modo, o conjunto de respostas candidatas a uma pergunta que verifiquem tal situação ganham vantagem em relação a outras onde apenas uma parte da entidade mencionada alvo é emparelhada. Ou seja, o sistema dá mais valor a uma dependência do tipo `PEOPLE_OK(Mário Soares, presidente de Portugal)` do que à dependência `PEOPLE_OK(Soares, presidente de Portugal)`, sendo a entidade alvo de exemplo `Mário Soares`.

Quanto a dependências compostas, em que o 3º argumento é um conceito auxiliar, normalmente um verbo ou adjectivo, acrescenta-se uma unidade à respectiva resposta candidata se tal conceito fizer também parte da lista de conceitos auxiliares contidos na pergunta. Por exemplo, para a questão "`Quando nasceu Álvaro Cunhal?`", o sistema prefere a dependência *DATE_OK(Álvaro Cunhal,10 de Novembro de 1913,Nascido)* a *DATE_OK(Álvaro Cunhal,10 de Novembro de 1913)* em que o conceito auxiliar patente na pergunta é *"nasceu"* e o que foi capturado na primeira dependência é *"Nascido"*. De salientar que, no caso dos verbos, apenas são comparados os quatro primeiros caracteres, com o intuito de ignorar diferentes conjugações verbais. Esta solução pode, contudo, dar origem a ambiguidades na comparação entre dois verbos que se distinguem entre si apenas a partir do quinto caracter. O único factor que se tinha em conta antes da implementação deste sistema de pesos - frequência das resposta candidatas - não foi esquecido, pelo que é acrescentado uma unidade à resposta candidata que ocorre mais vezes. Por fim, e como já foi acima referido, tem-se também em conta o valor de confiança da dependência, acrescentando-se uma unidade à resposta candidata se a dependência for do tipo OK.

São escolhidas pelo sistema no máximo três respostas distintas, com base no conjunto de todas as respostas candidatas e respectivos pesos. Caso não haja respostas candidatas nesta fase, isto é , se o sistema não detectou um padrão que originasse uma dependência adequada para responder à pergunta, são retornadas as três entidades mencionadas distintas mais frequente, do tipo esperado na resposta, nos textos seleccionados.

4.3 Técnica de extracção baseada em distâncias

Tendo como base fundamentos utilizados noutros sistemas de question-answering mais desenvolvidos, nomeadamente o da Priberam (Cassan et al., 2007), implementou-se uma nova abordagem de extracção de resposta baseada em distâncias entre conceitos presentes na pergunta e possíveis respostas. Inicialmente a técnica adoptada utilizava única e exclusivamente a Wikipédia como fonte de informação, sendo que mais tarde o trabalho foi expandido de maneira a que se pudesse usar a mesma técnica também para os textos do corpus jornalístico disponível.

Esta abordagem é constituída pelas seguintes fases:

- Determinar as páginas Wikipédia (ou artigos jornalísticos) a considerar,
- Determinar conceitos presentes nas perguntas e localizá-los no corpus,
- Seleccionar as frases a serem submetidas para a cadeia de Processamento de Língua Natural,
- Obter através da cadeia de PLN o conjunto de entidades mencionadas que possam servir de resposta à pergunta em questão e localizá-las no corpus,
- Calcular as distâncias, em número de palavras, para cada par conceito - resposta candidata,
- Escolher como resposta final a entidade mencionada que está mais perto de um dos conceitos presentes tanto na pergunta como no texto de suporte.

No caso da Wikipédia, torna-se mais fácil determinar que textos o sistema deve considerar, bastando captar o *target* da pergunta e interrogar a base de dados por páginas da Wikipédia cujo título contenha tal *target*. As páginas obtidas são guardadas num vector sobre o qual se executa um ciclo que tratará de cada uma delas. Na maior parte dos casos obtêm-se muitas páginas sem qualquer relação semântica com o *target* o que pode originar dois problemas:

- Consumo desnecessário de tempo na tentativa de procurar conceitos e respostas candidatas e calcular distâncias entre eles,

- O retorno por parte do sistema de uma resposta candidata incorrecta que eventualmente, mais tarde, possa ser escolhida como final.

Esta última hipótese é menos provável, visto que a página teria que conter pelo menos um conceito presente na pergunta bem como uma resposta candidata do tipo de que se está à espera na resposta.

No tratamento individual de cada página, é feita uma procura aos conceitos determinados na fase de análise e interpretação da pergunta. De seguida forma-se a frase a enviar para a cadeia de PLN, através de manipulação de cadeias de caracteres, que contêm tal conceito. Tendencialmente a frase em questão começa no primeiro caracter depois do primeiro ponto final antes do conceito e acaba no primeiro ponto final depois do conceito. Contudo esta abordagem exigiu que se ignorassem todos os caracteres '.' que não tinham a incumbência de finalizar uma frase. Por exemplo, se se submeter a pergunta "`Qual a profundidade da Fossa das Marianas?`" um dos excertos captados, e aquele que contém a resposta final, será:

```
A Fossa das Marianas é o local mais profundo dos oceanos,
atingindo 11.034 metros de  profundidade.
```

Se o primeiro '.' não fosse ignorado era somente considerado o excerto

```
A Fossa das Marianas é o local mais profundo
dos oceanos, atingindo 11.
```

Deste modo a resposta que o sistema poderia escolher no final seria "11" em vez de "11.034 metros".

A frase é submetida para a cadeia de PLN que tem como missão retornar as entidades mencionadas do tipo que se esta à espera na resposta, que, por sua vez, é determinado na fase de análise de interpretação da pergunta. Tais entidades extraídas através da cadeia de PLN são consideradas como respostas candidatas. São determinadas as posições em que se encontram os conceitos auxiliares e as respostas candidatas nos excertos em que elas ocorrem. Para cada par conceito - resposta candidata, contidos nesses excertos, calcula-se o módulo da diferença entre as duas posições, obtendo deste modo um conjunto de distâncias. Escolhe-se como resposta final, a resposta candidata que contenha a distância menor em relação a um dos conceitos considerados.

De salientar que os textos são previamente processados, retirando-se pontuações e guardando cada palavra sequencialmente num vector, cujo seu índice representa a posição da respectiva palavra

no texto. Uma das dificuldades detectadas na implementação desta técnica foi lidar com entidades mencionadas compostas e proceder à sua localização no texto de suporte. Visto que se está a considerar posições ordinais de palavras, a solução adoptada passou por fragmentar entidades mencionadas compostas. Se se considerar, por exemplo, como resposta candidata a entidade mencionada "Vila Nova de Gaia", para se determinar a sua posição no texto de suporte, não basta detectar a posição do referido vector que contenha a palavra "Vila" mas sim assegurar que as posições seguintes contenham as palavras "Nova", "de" e "Gaia". Convencionou-se que a posição de uma entidade mencionada composta é representada pelo índice da primeira palavra dessa entidade no vector de posições.

Com o intuito de minimizar a frequência de retorno de respostas incorrectas, determinou-se que só são aceites respostas com um limite configurável de distância máxima, isto é, se a resposta final escolhida por esta técnica de extracção exceder tal distância em relação a um dos conceitos considerados, o sistema redirecciona a pergunta para outra técnica de extracção de resposta. O fundamento para se ter tomado tal opção reside no facto de que quanto mais distante uma resposta candidata estiver de um conceito mais provável se torna não estarem relacionados entre si.

Passa-se agora a demonstrar todo o processo desencadeado por esta técnica com a pergunta de exemplo

```
Qual é o diâmetro de Ceres?
```

Na fase de análise e interpretação da pergunta é captado como target a entidade "Ceres" e como conceito auxiliar e fulcral para a determinação da resposta a entidade "diâmetro":

```
TARGET Ceres
ENTIDADES
AUXILIARES diâmetro target-type
```

O sistema redirecciona o tratamento desta pergunta para o *script* que trata das perguntas do tipo "Qual", que por sua vez, ao detectar a palavra "diâmetro" como um dos argumentos de *input* e percebendo de que se trata de uma medida, redirecciona para o *script* responsável pelo tratamento de perguntas do tipo "Quanto". Nesta fase o sistema já sabe a que tipo as respostas candidatas devem pertencer que, neste caso em particular, correspondem a quantidades, mais especificamente comprimento.

Já na execução da técnica em estudo são extraídas as páginas da Wikipédia que contenham a cadeia de caracteres "ceres" no título e para cada uma delas localiza-se a palavra "diâmetro". Na página com o título "Ceres_(planeta_anão)" é encontrada uma ocorrência do conceito auxiliar "diâmetro" e é recolhido o seguinte excerto que será posteriormente submetido para a cadeia de PLN:

```
Ceres tem um  diâmetro de cerca de 950 km e é o corpo mais maciço
 dessa região do sistema solar, contendo cerca de um terço do total
 da massa da cintura
```

A cadeia de PLN devolve a única entidade mencionada do tipo *Quantidade*, presente na frase - 950 km - que é considerada como resposta candidata. As posições na frase de suporte da palavra "diâmetro" e "950 km" são 3 e 7 respectivamente pelo que a distância calculada é de 4 unidades. Como foi a única resposta candidata considerada, de todas as páginas obtidas, e como a distância entre ela e o conceito auxiliar não ultrapassa o limite estipulado, é retornada pelo sistema como resposta final.

A grande limitação desta estratégia surge quando uma pergunta contém múltiplos conceitos auxiliares. O grande problema reside no facto de ser impossível o sistema perceber qual, de todos os conceitos, aquele que tem mais impacto na pergunta e, consequente e hipoteticamente, o que é mais determinante para a extracção da resposta. Na pergunta *"Por que estado foi eleito o senador Barack Obama?"* coabitam, para além do *target "Barack Obama"*, várias entidades que podem ser consideradas como conceitos auxiliares - o verbo *"eleito"*, o cargo *"senador"* e *"estado"*. Tais entidades são consideradas pelo sistema, aumentando o número de partes dos textos de suporte a extrair e distâncias a calcular e levando, deste modo, a um significativo aumento do peso computacional. Mais grave do que isso, é o aumento da probabilidade do sistema retornar uma resposta incorrecta quanto maior for o número de conceitos auxiliares. O facto de estes serem tratados com o mesmo grau de relevância pode originar a que o sistema recolha uma resposta candidata incorrecta por esta estar perto de um conceito pouco importante também presente na pergunta.

Esta técnica apresenta ainda vantagens a nível de desempenho devido a de dois factores. O primeiro está relacionado com a interrogação feita à base de dados, que não sendo feita ao texto integral das páginas da Wikipédia, mas sim somente ao título dessas páginas, possibilita o retorno mais célere dos resultados. Para além disso, e ao contrário do que é feito noutras técnicas de extracção (nomeadamente a técnica de Força Bruta descrita em 2.3.2 e 4.2), só são submetidos para a cadeia de PLN pequenos excertos previamente seleccionados, o que agiliza significativamente o seu processamento, quando comparado com uma possível solução alternativa que passasse por submeter toda uma página da Wikipédia ou um determinado artigo do corpus jornalístico.

4.3.1 Extracção de informação através de expressões regulares

Outro aspecto positivo desta abordagem consiste na exploração da forma bem organizada e uniforme da informação disponível na Wikipédia, possibilitando a extracção imediata, sem recorrer a cálculos de distâncias, de informações como, por exemplo, datas e locais de nascimento e de determinados acontecimento. No Wikipédia, para perguntas do tipo "*Quando nacseu/morreu X?*", o sistema extrai a página com o título *X* que, usualmente se inicia com o seguinte padrão:

```
X (local_de_nascimento, data_de_nascimento –
local_de_falecimento, data_de_falecimento) foi...
```

Se tal padrão for detectado, consegue-se, através da manipulação de cadeia de caracteres, obter a informação pretendida sem se recorrer a nenhum cálculo de distâncias. Por exemplo, para a pergunta "*Quando nasceu Ayrton Senna?*" o artigo da Wikipédia recolhido começa deste modo:

```
Ayrton Senna da Silva (São Paulo, 21 de março de 1960
| Bolonha, Itália, 1 de maio de 1994)
foi um piloto brasileiro de Fórmula 1...
```

Visto que na pergunta se encontra o vocábulo *nasceu*, cujo lema é *nascer*, o sistema vai extrair a primeira data que encontra no referido padrão - *21 de março de 1960*. Tendo em conta que as datas aparecem na Wikipédia num formato normalizado, achou-se conveniente extraí-las a partir da manipulação de cadeia de caracteres. Outra solução seria submeter a primeira frase do artigo Wikipédia à cadeia de PLN e extrair as datas que lá se encontrariam. Contudo, tal solução implicaria um peso computacional extra que não traria resultados significativos. Para além disso, se a primeira frase emparelha com o padrão acima transcrito, tem-se a certeza que a primeira data corresponde ao nascimento da entidade em causa e a segunda à sua morte. Se se submeter cada frase que contenha datas, do artigo Wikipédia, à técnica de extracção baseada em distâncias, pode-se dar o caso de que outras datas sejam consideradas e, se se encontrarem mais perto dos conceitos auxiliares, serem as escolhidas para resposta final. Posto isto, uma expressão regular para captar datas no formato usual apresentado na Wikipédia é:

```
(\w+\s)(de\s)(\w+\s|março\s)(de\s)(\d\d\d\d)
```

Para perguntas que restringem o domínio da resposta do tipo data as suas respostas sofrem também uma manipulação de cadeia de caracteres, explicadas na tabela 4.1 que contém, para cada uma das três restrições típicas, explicitadas com três exemplos, que se fazem em datas, a parte da expressão regular que nos interessa extrair da data completa. Tal manipulação é feita independentemente da data ter sido extraída através de extracção baseada em distâncias ou pela detecção, na primeira frase do artigo Wikipédia, do padrão previamente transcrito.

Restrição	Exp. Reg.	Resposta
Em que **ano** nasceu Ayrton Senna?	`(\d\d\d\d)`	1960
Em que **mês** nasceu Ayrton Senna?	`(\w+\s|março\s)(de\s)(\d\d\d\d)`	Março de 1960
Em que **dia** nasceu Ayrton Senna?	`expr. completa`	21 de Março de 1960

Tabela 4.1: Restrições impostas pelas perguntas de respostas com datas

4.4 Módulo de Validação da Resposta

Como já foi realçado, este trabalho visa também a elaboração dos primeiros passos de um sub-sistema de validação de resposta a integrar com o sistema principal. Tal módulo tem dois grandes objectivos em perspectiva:

- Evitar que o sistema retorne uma resposta semanticamente errada, isto é, uma reposta de uma classe diferente da esperada. Por exemplo, evitar que para perguntas da categoria "Onde" se obtenham respostas do tipo *PEOPLE*.
- Determinar o nível de fundamentação da resposta, com base no respectivo texto de suporte.

Para alcançar estes dois objectivos foram aproveitadas e tidas em consideração várias abordagens descritas e analisadas em 3.3. Tratam-se de técnicas comuns em sistemas de validação de resposta, com participação no AVE, e que serviram de base para a implementação deste módulo.

4.4.1 Visão Global

O módulo de validação da resposta implementado, cuja visão global se encontra esquematizada na Figura 4.1, é composto por dois sub componentes, que procuram dar resposta aos dois objectivos acima referidos:

- validação do tipo da resposta,
- verificação da fundamentação da resposta.

Do ponto de vista prático o módulo em questão pode ser visto como um objecto que contém os seguintes elementos:

- a resposta candidata R,
- O conjunto das entidades mencionadas e vocábulos auxiliares considerados relevantes contidas na pergunta, EM,

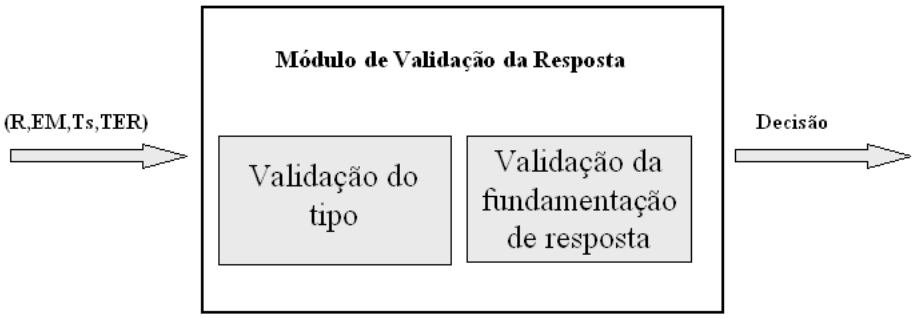

Figura 4.1: Visão global do módulo de validação da resposta.

- o texto que suporta R, T_s,
- o tipo esperado na resposta, TER, com base na informação recolhida na fase de análise e interpretação da pergunta, ou seja, a classe a que a resposta pertence de entre as seguintes:
  - LOCATION; PEOPLE; TITLE; DATE; QUANT; EVENT
- um conceito auxiliar - *target-type* - que possa estar contido na pergunta e que especifique um sub-tipo da resposta, restringido o domínio da mesma.

De salientar que este último argumento é opcional, isto é, pode haver perguntas onde o *target-type* é vazio. Por exemplo, para a pergunta *"Onde nasceu Álvaro Cunhal"*, nenhum *target-type* é detectado, sendo o tipo esperado na resposta a própria classe LOCATION. Tal já não se sucede com a pergunta *"Em que cidade nasceu Álvaro Cunhal"*. O sistema marca o vocábulo *cidade* como *target-type*. É feito um redireccionamento do vocábulo captado para o respectivo sub-tipo que o XIP considera. Neste caso o tipo esperado na resposta é *CITY*, que é uma sub-classe da classe LOCATION. Só as classes *LOCATION*, *QUANT* e *DATE* contêm sub-classes captadas nos *target-type* contidos na pergunta. A Figura 4.2 exibe os vários sub-tipos pertencentes a estas classes.

Com base nos argumentos recebidos aquando da construção de uma instância do objecto em causa, é feita a validação do tipo da resposta bem como a verificação de que tal resposta se encontra devidamente fundamentada. Tais funcionalidades sao descritas mais aprofundadamente em 4.4.2 e em 4.4.3.

4.4.2 Validação do tipo da resposta

De modo idêntico ao que é feito no sistema da UAIC (Iftene & Balahur-Dobrescu, 2007, 2008) analisado em 3.3.3, o método responsável pela validação do tipo da resposta tem como argumentos a resposta candidata, a sua classe e possíveis conceitos - *target-type* - que possam restringir o domínio da resposta a uma determinada sub-classe. A resposta candidata é submetida à cadeia de PLN e são extraídos e

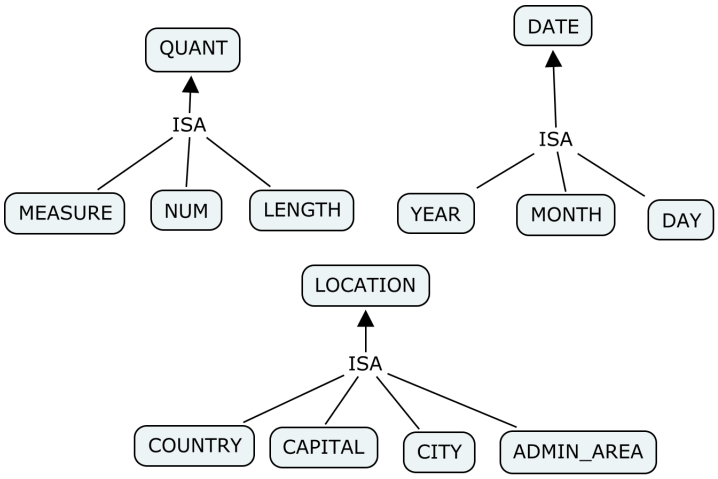

Figura 4.2: Classes e sub-classes de tipos de resposta

guardados todos os traços que foram associados à resposta em questão. A partir da classe da resposta e do *target-type* é calculado o tipo específico esperado na resposta. Este tipo corresponde a um traço do conjunto de traços considerados pela cadeia de PLN. Posto isto, torna-se necessário efectuar o mapeamento de vocábulos captados na pergunta e tratados como *target-type* para o respectivo traço da cadeia de PLN. Por exemplo, se o sistema apanhar na pergunta *target-types* como "região", "estado" ou "concelho" tem que ser feito um redirecionamento para o traço *ADMIN_AREA*, que é visto como um sub-tipo de *LOCATION* (ver 4.2). Se não houver *target-type* associado à pergunta, o tipo esperado na resposta será o mesmo que a sua classe. Ou seja, para uma pergunta do tipo "Onde" sem *target-types* associados, o tipo da entidade mencionada esperada na resposta é *LOCATION*.

Depois de obtido o tipo específico esperado na resposta, o sistema verifica se o mesmo está contido na lista de traços identificados na resposta candidata. Caso isso aconteça, é atribuída a pontuação de uma (1,0) unidade. Se, por outro lado, o tipo da resposta candidata for uma super classe do tipo específico de que se espera na resposta é atribuída a pontuação de 0,5. Caso nenhum dos casos se verifique, é atribuída a pontuação de 0. Considere-se a seguinte pergunta: "`Em que distrito fica Chaves?`" e a resposta candidata "`Distrito de Vila Real`". O sistema captou o vocábulo "*distrito*" como *target-type* da pergunta e como consequência mapeou-o para o traço *ADMIN_AREA*, que se torna o tipo específico esperado na resposta. Ao se submeter a resposta na cadeia de PLN obtém-se a respectiva lista de traços transcrita parcialmente de seguida:

```
"Traços: ...START LAST FIRST CITY ADMIN_AREA LOCATION NOUN END... "
```

Como o tipo esperado na resposta se encontra na lista de traços associados à resposta candidata, o sistema de validação do tipo da resposta retorna a pontuação máxima de 1,0. De notar, que caso

a pergunta fosse somente "*Onde fica Chaves?*" obteria-se o mesmo resultado para a mesma resposta candidata, visto que não havia nenhum *target-type* definido, pelo que o tipo esperado na resposta seria *LOCATION* que também se encontra na lista de traços acima transcrita.

Este sistema tem algumas limitações, estando dependente do correcto comportamento da cadeia de PLN. Se na pergunta de exemplo anterior o XIP não associasse, à resposta candidata, o tipo *ADMIN_AREA* a pontuação obtida seria de 0,5 e de 0 caso não associasse o tipo *ADMIN_AREA* nem o tipo *LOCATION*.

4.4.3 Verificação da fundamentação da resposta

Para verificar se a resposta candidata é minimamente fundamentada pelo respectivo texto de suporte, o sistema formula uma hipótese, que consiste no conjunto resultante da união dos vocábulos das entidades mencionadas e de outros conceitos auxiliares patentes na pergunta, como o *target-type*, e os vocábulos da resposta. De salientar que, ao integrar este módulo de validação de resposta com o sistema de QA desenvolvido, tem se a certeza que pelo menos um vocábulo da resposta e um das entidades mencionadas da pergunta estão contidas no texto de suporte, pelo que a verificação da fundamentação de resposta usada no referido sistema de QA irá somente quantificar o grau de fundamentação de uma resposta candidata com base no respectivo texto de suporte. Este sub-componente do módulo de validação de resposta visa calcular determinadas métricas que consoante o seu valor conduzem, ou não, à validação da resposta candidata, e que passam a ser descritas nos próximos sub capítulos.

4.4.3.1 Cálculo do rácio entre conceitos patentes na pergunta e na resposta

Numa primeira etapa, é feita a contagem dos vocábulos da hipótese que também estão contidos no texto de suporte e atribuída uma pontuação segundo a seguinte fórmula:

$$\frac{(N_{em} + N_{resp})}{N_{hipotese}}$$

em que N_{em} é o número de os vocábulos da(s) entidade(s) mencionada(s) contidas na pergunta que surgem também no texto de suporte, N_{resp} o número dos vocábulos da resposta também presentes no mesmo texto e $N_{hipotese}$ o número de elementos da lista de vocábulos da hipótese previamente formulada. Caso N_{em} ou N_{resp} sejam zero é sinal de que nenhuma palavra das entidades mencionadas da pergunta, ou nenhuma palavra da resposta candidata, se encontra no texto de suporte, pelo que se decidiu admitir que tal resposta não está minimamente fundamentada, atribuindo-se o valor 0 para estes casos. Como já foi referido, tal situação não pode acontecer no sistema de QA desenvolvido, pois este garante que pelo menos um vocábulo das respostas candidatas e das entidades contidas na pergunta se encontra também presente no texto de suporte em questão.

Mostra-se agora, com um exemplo, o funcionamento deste componente, considerando a pergunta *"Onde nasceu Barack Obama?"*, a resposta candidata *"Havaí"* e o seguinte texto de suporte:

```
"Obama é o único senador com ascendência africana na actual legislatura.
Obama  nasceu no Havaí quando seu pai, um queniano, e sua mãe,
Ann Dunham de Wichita Kansas, estudavam na Universidade do Havaí em Manoa"
```

Com este *input*, o sistema produz a lista de vocábulos a partir da entidade mencionada na pergunta e a resposta candidata, obtendo, deste modo a hipótese *"Barack Obama Havaí"*. O sistema encontra dois vocábulos contidos tanto na hipótese como no texto de suporte - *Obama* e *Havaí* - sendo que o primeiro foi obtido através do conjunto dos vocábulos das entidades da pergunta e o segundo na resposta, pelo o que N_{em} e N_{resp} têm, neste exemplo, o valor de uma unidade. Assim sendo, o sistema aceita a fundamentação da resposta atribuindo a pontuação 0,66, de acordo com a fórmula anteriormente referida. Caso se se considerasse um texto de suporte onde também, para além de *Obama* e *Havaí*, estivesse contida a palavra *Barack*, obter-se-ia, neste componente de verificação da fundamentação da resposta, a pontuação máxima de 1 valor.

Outro aspecto que importa realçar reside no facto de, em muitos casos, serem apresentadas grafias distintas para referir a mesma entidade mencionada. Tal situação pode levar a que o sistema não admita que uma determinada entidade seja fundamentada no texto de suporte, visto que se encontra escrita de uma maneira diferente. Com o objectivo de contornar este problema, especialmente para os casos onde as diferenças entre as grafias são minímas, e à semelhança do que é feito em (Castillo, 2008; Rodrigo et al., 2007), o sistema calcula a distância de Levenshtein entre cada vocábulo da hipótese e cada um dos vocábulos do texto de suporte. Considera-se que duas palavras referem-se à mesma entidade se a distância de Levenshtein entre elas não exceder os 20%. Ou seja, por cada cinco caracteres de uma palavra, o sistema admite uma das três operações consideradas no referido algoritmo - inserção, eliminação e substituição. Deste modo, e aproveitando o exemplo escolhido e descrito no parágrafo anterior, a pontuação do componente de verificação da fundamentação de resposta não se alteraria caso a resposta candidata fosse Havai em detrimento de Havaí, visto que apenas é feita uma substituição no último caracter de uma palavra com cinco letras, pelo que a distância de Levenshtein entre estes dois vocábulos difere somente em 1/5 - 20%. Visto que o cálculo da diferença da distância de Levenshtein entre os dois vocábulos se assenta no número de caracteres do maior vocábulo, tem que se ter o cuidado de se assegurar que os vocábulos não estão codificados no formato *Unicode*, pois tal codificação conduz a resultados incorrectos na contagem de letras em vocábulos com caracteres especiais.

4.4.3.2 Cálculo das distâncias entre resposta candidata e entidades da pergunta

Com base no que é feito em vários sistemas de QA, nomeadamente o da Priberam (Cassan et al., 2007), a distância, num determinado texto de suporte, entre um conceito *pivot*, que para estes casos se tratará sempre da resposta candidata, e cada um dos conceitos auxiliares, nomeadamente entidades mencionadas, *target-types* ou verbos, pode ser vista como um importante factor de validação de uma resposta candidata. Quanto mais perto um conceito, ou conjunto de conceitos, estiver da resposta candidata, maior é a qualidade atribuída ao texto de suporte que fundamente tal resposta. Posto isto, o módulo de verificação da fundamentação da resposta, contempla também o cálculo de todas as distâncias entre a resposta candidata e cada uma das entidades mencionadas e *target-types*, presentes tanto na pergunta como no texto de suporte, reutilizando o método de cálculo das distancias implementado na técnica de extracção de resposta baseada em distâncias 4.3. Considerando a pergunta de exemplo *Em que estado nasceu Barack Obama?*, a resposta candidata *Havaí* e o texto de suporte "`... Obama nasceu no estado de Havaí quando seu pai...`", é gerado um vector de distâncias onde são explicitamente guardadas as distâncias entre os pares (Havaí, Obama) e (Havaí, estado) - cinco e duas unidades respectivamente.

4.4.4 Cálculo da pontuação final

Com base no somatório das distâncias calculadas, no comprimento do texto de suporte - número de palavras - e no rácio previamente calculado e explicado em 4.4.3.1, é obtida uma pontuação final que atribui um grau de qualidade à fundamentação da resposta candidata, considerando o seu respectivo texto de suporte. A referida pontuação é obtida através da formula:

$$\frac{R * L}{\sum_{i=0}^{n} dist(resp, concept_i)}$$

em que R é o rácio calculado de acordo com o método descrito em 4.4.3.1, L o número de palavras, n o número total de conceitos da pergunta tidos em conta e dist(resp, concept$_i$) a função que retorna a distância entre a resposta candidata *resp* e o conceito *concept*$_i$. Com base nesta fórmula, quanto menor o somatório das distâncias e maior o rácio entre o numero de palavras presentes na pergunta e, também, no texto de suporte e o número de palavras presentes apenas na pergunta, melhor é a qualidade de fundamentação da resposta candidata em questão.

Como alternativa à abordagem previamente descrita, foi adoptada outra métrica que se baseia na distância média entre conceitos - entidades e *target-types* - da pergunta e a resposta candidata, que é

dada por:

$$\frac{\sum_{i=0}^{n} dist(resp, concept_i)}{N}$$

em que N é o número de vocábulos da lista de conceitos considerados na pergunta que também se encontram presentes no texto de suporte. Ao dividir-se o número de palavras do texto de suporte pela distância média entre conceitos obtém-se o resultado final deste componente de validação da fundamentação da resposta, seguindo esta abordagem alternativa. De forma análoga ao cálculo explicado no parágrafo anterior, premeiam-se textos de suporte que contenham um número elevado de vocábulos contidos na pergunta e que apresentam pequenas distâncias entre tais vocábulos e a resposta candidata.

Não obstante o tipo de abordagem escolhido para o cálculo da métrica a ser considerada, o resultado parcial obtido no sub componente de validação da fundamentação da resposta é multiplicado pela pontuação atribuída à resposta candidata em avaliação pelo sub componente de validação do tipo da resposta descrito em 4.4.2. O produto obtido é a pontuação final, atribuída pelo módulo de validação da resposta, para cada triplo *(pergunta, resposta candidata, texto de suporte)*. Cabe, por fim, ao módulo em questão validar a resposta candidata se tal pontuação exceder um *threshold* mínimo - 3.00 - previamente estipulado, com base em análises empíricas, ou rejeitar a resposta caso contrário.

5 Avaliação

5.1 Introdução

O projecto Clefmania 2 participou no fórum CLEF em Maio de 2008, pelo que foram aproveitadas as metodologias de avaliação impostas por este fórum. Tal como em anos anteriores, a avaliação foi feita com base na submissão de 200 perguntas de três diferentes tipos:

- Definition (D) - pergunta de definição como por exemplo "*Quem é Mário Soares?*" ou "*O que são os forcados?*".
- Factoid (F) - perguntas baseadas em factos, cuja resposta é, tipicamente, uma entidade mencionada de um tipo determinado depois de realizada a interpretação da pergunta. "*Quantos ossos têm a face?*" e "*Onde fica Saint-Exupéry?*" são dois exemplos de perguntas do tipo *Factoid*.
- List (L) - perguntas que têm como resposta um determinado número de items. Por exemplo, "*Quais são as regiões da Bélgica?*" é uma pergunta do tipo *List*.

Para além desta classificação de tipos de perguntas, ainda se consideram questões do tipo NIL que se caracterizam por não terem qualquer resposta. São exemplos disso perguntas como "*Que mar banha Braga?*" ou "*De que país Nova York é capital?*".

Tal como na edição anterior criaram-se grupos de perguntas de forma a conter figuras de estilo como a anáfora e elipse, como por exemplo:

```
group_id="2662": Quem foi o último rei de Portugal?
group_id="2662": Em que período foi ele rei?
group_id="2662": Em que barco ele embarcou para o exílio?
```

A avaliação a realizar ao sistema de QA, será feita com base no tipo de pergunta de forma a poder verificar para quais o sistema apresenta melhores e piores resultados. O método de avaliação usado na edição de 2008 do CLEF não difere em nada do da edição anterior de 2007, pelo que são considerados quatro tipos de resultados que uma resposta submetida pelo sistema de QA pode ter:

- R(Right) - se a resposta for correcta e devidamente fundamentada pelo texto de suporte,

- W(Wrong) - se a resposta for errada,
- X(ineXact) - se a resposta contiver mais ou menos informação do que aquele pretendida pela pergunta,
- U(Unsupported) - se a resposta não estiver contida no texto de suporte a ela associado ou se o *id* do documento de suporte estiver errado ou em falta.

A métrica considerada como primordial para avaliação de sistemas QA é a precisão (*accuracy*) e consiste na média da uma função SCORE(q) para o número de perguntas de um conjunto a avaliar, em que:

$$SCORE(q) = \begin{cases} 1, & \text{se q for R;} \\ 0, & \text{c.c.} \end{cases}$$

Posto isto, tem-se:

$$precisão = \sum_{q=0}^{Q} \frac{SCORE(q)}{Q}$$

em que, tipicamente no fórum CLEF, Q=200 - número total de perguntas submetidas e avaliadas por cada submissão.

Não obstante o facto do módulo de validação de resposta, integrado no sistema QA@L^{2}F, ter sido implementado já depois da participação do mesmo sistema na edição de 2008 do fórum CLEF, efectuar-se-á a sua avaliação com base em métricas consideradas no âmbito do *Answer Validation Exercize* - AVE - de 2008. De acordo com as normas do AVE, é atribuído um dos seguintes resultados a cada resposta candidata submetida a um sistema de validação da resposta:

- VALIDATED, se a resposta r à pergunta p é provada pelo texto de suporte t;
- SELECTED, para a melhor resposta r do conjunto das respostas marcadas como VALIDATED;
- REJECTED, se a resposta r à pergunta p não é provada pelo texto de suporte t.

As métricas usadas para avaliação de sistemas de validação de resposta têm em conta tal espectro de possíveis resultados e obtêm-se a partir das seguintes fórmulas:

$$precision = \frac{\#respostas\_correctamente\_selecionadas\_ou\_validadas}{\#respostas\_selecionadas\_ou\_validadas}$$

$$recall = \frac{\#respostas\_correctamente\_selecionadas\_ou\_validadas}{\#respostas\_certas}$$

$$f - measure = \frac{2 * recall * precision}{recall + precision}$$

$$qa - accuracy = \frac{\#respostas\_correctamente\_selecionadas}{\#perguntas}$$

As três primeiras medidas servem para quantificar a habilidade do sistema em detectar se existe fundamentação suficiente para aceitar uma resposta, enquanto que a última permite obter uma comparação entre a eficácia de sistemas de validação e sistemas de QA correspondentes.

5.2 Avaliação do sistema de QA

Como é usual nas anteriores edições do CLEF são avaliadas duas submissões com os resultados obtidos com o mesmo conjunto de teste. O conjunto de teste contém 200 perguntas que podem ser consultadas no anexo A e se distribuem pelas várias categorias da seguinte maneira:

	N.º Perguntas	Percentagem
Factoid	162	81.0%
Definition	28	14.0%
List	10	5.0%
R. Temp	16	8.0%
NIL	-	-
Total	**200**	100%

Tabela 5.1: Categorias de perguntas e respectivo peso no conjunto de teste

De salientar que perguntas com restrições temporais e perguntas cuja resposta é *NIL* não são categorias de perguntas mas sim características pelo que estão contidas numas das três categorias consideradas.

	N.º de respostas	**% de respostas**	**N.º de respostas**	**% de respostas**
Right	38	19.0%	41	20.5%
Wrong	149	74.5%	148	74.0%
ineXact	6	3.0%	5	2.5%
Unsupported	7	3.5%	6	3.0%
Total	200	100.00	200	100.00

Tabela 5.2: Resultados obtidos pelo QA@L^{2}F na 1ª e 2ª submissão.

A Tabela 5.2 mostra os resultados obtidos nas duas submissões feitas no âmbito do CLEF 2008. São observadas pequenas melhorias nos resultados ao comparar-se a segunda com a primeira submissão. Tal facto é justificado pela correcção de pequenos erros no sistema que impediam o retorno do texto ou frase de suporte completo, levando a que o resultado de algumas respostas fosse *Unsupported*, ou, noutros casos, o retorno da resposta completa de perguntas da categoria *Definition*, que produzia o resultado *ineXact* para as respectivas respostas.

A Tabela 5.3 mostra que perguntas obtiveram resultados distintos entre a 1ª e a 2ª submissão bem como as causas que explicam tal diferença de comportamento.

Pergunta	Resp.I	Res.I	Resp.II	Res.II	Causa
0021	*NIL*	W	12 de Agosto de 1955	R	A informação escondida pela a anáfora não estava a ser recuperada na fase de interpretação da pergunta
0027	um pedagogo	X	um pedagogo, editor e enciclopedista francês	R	A técnica de emparelhamento linguístico considerava o cadeia de caracteres da resposta até à 1ª ',' em vez de considerá-la até ao 1º '.'
0053	uma sopa típica do Alentejo – ao contrário da maioria das sopas	X	uma sopa típica do Alentejo – ao contrário da maioria das sopas, esta não é cozinhada, mas basicamente pão em água quente temperada	R	*idem*
0106	comprimento de cerca de 400 km	U	comprimento de cerca de 400 km	X	O texto de suporte retornado era igual à resposta pois acabava na 1ª ',' em vez de acabar no 1º '.'

Tabela 5.3: Diferenças de resultados entre a 1ª e 2ª submissão

Numa avaliação global ao sistema e comparando-o com a sua participação no CLEF 2007, pode-se constatar, com auxílio dos resultados obtidos e transcritos nas tabelas 5.5 e 5.4, que:

- apesar de uma significativa melhoria nas respostas às perguntas *Factoid* - de 5.03% em 2007 para 13.58% em 2008 - a taxa de precisão ainda se encontra aquém da esperada,
- houve um ligeiro aumento na precisão para perguntas *Definition* de 58% para $60,71\%$,
- uma resposta certa do tipo *List* e outra inexacta foram respondidas correctamente o que implica um melhoramento - de 0% para 10.0% - maioritariamente explicado pelo o facto de que, aquando a participação na edição de 2007 do CLEF, a fase de análise e interpretação de perguntas ainda não tratar perguntas desta categoria.
- quatro perguntas *Factoid* foram avaliadas como *Unsupported* pois as respostas foram retiradas de tabelas da Wikipédia e a sua fundamentação foi ignorada. No total o sistema retornou na 2ª submissão do CLEF 2008 52 respostas não erradas (*Right* + *IneXact* + *Unsupported*) que corresponde a 26% das perguntas do conjunto de teste e a um aumento de 10% em relação à 2ª submissão feita no ano anterior.

	Factoid	*Definition*	*List*	*Temp. Restricted*	*NIL*
Right	8	18	0	1	11
Wrong	150	8	10	18	141
ineXact	0	4	0	0	0
Unsupported	1	1	0	0	0
Total	159	31	10	19	152
Precisão	5.03%	58.06%	0.00%	5.26%	7.24%

Tabela 5.4: Resultados detalhados obtidos pelo QA@L^{2}F na 2ª submissão do CLEF 2007.

	Factoid	*Definition*	*List*	*Temp. Restricted*	*NIL*
Right	22	17	2	1	9
Wrong	132	10	13	11	-
ineXact	3	1	1	0	-
Unsupported	5	0	0	1	-
Total	162	28	10	16	-
Precisão	$22/162 \simeq 13.58\%$	$17/28 \simeq 60,71\%$	$1/10 = 10.00\%$	$1/16 \simeq 6.25\%$	-

Tabela 5.5: Resultados detalhados obtidos pelo QA@L^{2}F na 2ª submissão do CLEF 2008.

De salientar ainda que para a melhoria dos resultados, contribuiu também o facto de haver na fase de interpretação e análise da pergunta o tratamento de anáforas, do qual este trabalho se incidiu, que permitiu responder correctamente a quatro perguntas (ver anexo perguntas com: id=0003, id=0068, id=0157 e id=0191) que no CLEF de 2007 seriam impossíveis de serem respondidas.

Sugere-se agora uma avaliação das várias técnicas de extracção de resposta utilizadas, contabilizando o número de respostas correctas que cada uma delas conseguiu retornar na segunda submissão efectuada no âmbito do forum de avaliação CLEF 2008. As técnicas de extracção em causa são:

- Emparelhamento de Padrões Linguísticos - EPL,
- Reordenação de Formulações Linguísticas - RFL,
- Técnica de Extracção Baseada em Distâncias - TEBD,
- Força Bruta com pós-processamento de Língua Natural - FB.

A Tabela 5.6 expõe o número de respostas que cada uma das referidas técnicas conseguiu responder com *Right IneXact* e *Unsupported* e respectivo peso percentual no total de respostas retornadas pelo sistema na 2ª submissão feita no âmbito do fórum de avaliação CLEF 2008.

Na fase de extracção da resposta o sistema caracteriza-se por executar numa determinada sequência, previamente estabelecida consoante o resultado da fase de análise e interpretação da pergunta, um conjunto de técnicas de extracção de resposta. Para respostas *NIL*, à medida que as várias

Técnica de Extracção	*Right*	*IneXact*	*Unsupported*
EPL	3 (7.3%)	0 (0.0%)	0 (0.0%)
RFL	14 (34.1%)	2 (40.0%)	4 (66.6%)
TEBD	12 (29.3%)	2 (40.0%)	1 (16.6%)
FB	12 (29.3%)	1 (20.0%)	1 (16.6%)
Total	41 (100%)	5 (100%)	6 (100%)

Tabela 5.6: Performance das técnicas de extracção.

técnicas não extraem qualquer resposta, o sistema acaba sempre por executar a técnica de Força Bruta que, por ser o último recurso, há-de sempre dar uma resposta mesmo para os casos em que esta seja, correcta ou incorrectamente, *NIL*. Por essa razão, nove das doze respostas correctas extraídas pela Técnica de Força Bruta são respostas certas com o valor *NIL*. Outro aspecto a realçar consiste no baixo número de respostas correctas extraídas pela técnica de Emparelhamento e de Padrões Linguísticos que pode ser explicado por dois motivos:

- a informação disponível para esta técnica foi extraída a partir do pré-processamento de apenas 30% do corpus,
- na altura do último pré processamento realizado ainda não estavam implementados novos padrões de detecção de relações entre entidades mencionadas.

Outro dado que a tabela mostra diz respeito à nova técnica desenvolvida no lectivo 2007/2008 - Técnica de Extracção Baseada em Distâncias (4.3) - que respondeu a 30% do total das respostas correctas e mostrou ser a grande responsável pela significativa melhoria do sistema para perguntas do tipo *Factoid*. Os resultados obtidos com esta técnica de extracção poderiam ter sido melhores se não se verificassem alguns casos em que, apesar desta técnica ter adoptado o comportamento pretendido, a resposta correcta não tenha sido retornada devido a falhas de etiquetação de entidades mencionadas, imprescindíveis para o retorno da resposta candidata. Por exemplo, para a pergunta "`Quantos jogadores tem uma equipa de voleibol?`" (id=0140) o sistema não produz qualquer resposta. Contudo, ao se estudar mais aprofundadamente o comportamento do sistema, verifica-se que na técnica de extracção baseada em distâncias é escolhida a seguinte frase, retirada de uma página da Wikipédia, a ser submetida para a cadeia de PLN:

```
"Voleibol é um desporto praticado numa quadra
dividida em dois por uma rede, por duas equipas
de seis jogadores cada."
```

Sendo a pergunta do tipo "Quanto", são apenas consideradas como possíveis respostas candidatas as entidades mencionadas marcadas pelo XIP com a etiqueta *QUANT*. O facto da resposta candidata -

"*seis*" - estar escrita por extenso faz com que o XIP não a reconheça como uma quantidade e não faz a associação necessária entre o vocábulo em causa e a etiqueta *QUANT*.

De referir também um pequeno pormenor que impediu a resposta correcta de pelo menos uma pergunta no conjunto de teste em questão. Tal pormenor está relacionado com as *query's* que se fazem com base títulos de páginas Wikipédia. Para títulos compostos por mais de que um vocábulo, é necessário considerar duas alternativas possíveis:

- vocábulos separados por espaços (exemplo: Mário Soares),
- vocábulos separados por "_"(*underscore*) (exemplo: Mário_Soares).

O sistema não respondeu à pergunta "*Quem foi Carl Barks?*" exactamente porque tentou extrair a pagina da Wikipédia com o titulo "*Carl Barks*", não obtendo nenhum resultado e ignorando a alternativa "*Carl_Barks*". Como as outras técnicas de extracção também não produziram qualquer resultado a resposta final foi *NIL*. Ao considerar-se como título a alternativa "*Carl_Barks*", obtém-se a página pretendida. A técnica de Reordenação de Formulações Linguísticas comporta-se da maneira adequada, retornando a seguinte resposta baseada no respectivo texto de suporte:

```
Resposta: ''um famoso ilustrador dos estúdios Disney
e criador de arte sequencial...''

Suporte: ''Carl Barks (27 de Março de 1901–25 de Agosto de 2000) foi um
famoso ilustrador dos estúdios Disney e criador
de arte sequencial, responsável pela invenção de Patópolis
e muitos de seus habitantes...''
```

5.3 Avaliação do módulo de validação da resposta

O módulo de validação da resposta foi implementado já depois da participação do sistema QA@L^2F na edição de 2008 do CLEF pelo que não foi possível submeter o referido módulo a uma avaliação oficial no âmbito do AVE 2008. Apesar disso, decide fazer-se nesta secção uma avaliação do módulo implementado, tendo em conta as métricas usadas no AVE e descritas em 5.1 e reaproveitando o conjunto de teste do CLEF 2008. Deste modo, para o cálculo da *qa_accuracy* são contabilizadas as respostas correctamente seleccionadas pelo módulo de validação num total de 73 perguntas: as 32 respostas *Right* e 34 respostas *Wrong* que não são *NIL*, as cinco *ineXactas* e duas das seis *Unsupported*, visto que em quatro delas o suporte se referia a tabelas da Wikipédia e foi totalmente ignorado. Para o cálculo do *recall*, são contabilizadas as respostas correctamente seleccionadas num total de 32 perguntas que correspondem

ao número de respostas certas do sistema QA@L^2F, excluindo as correctas com o valor *NIL*, na 2ª submissão efectuada no CLEF 2008. Finalmente, para o cálculo da *precision* determina-se a relação entre o número de respostas correctamente seleccionadas e o total de respostas seleccionadas de entre as 73 perguntas submetidas a avaliação.

A Tabela 5.7 mostra o número de respostas validadas pelo módulo a partir das duas métricas utilizadas e explicadas na secção 4.4.3, no conjunto de teste formado por triplos (pergunta, resposta candidata, texto de suporte) a partir dos resultados retornados pelo sistema de QA às 73 perguntas das 200 do CLEF 2008 que contemplam os requisitos já referidos. As validações estão agrupadas em quatro grupos que correspondem aos 4 estados possíveis que uma resposta pode ter no fórum de avaliação CLEF - *Right*, *ineXact*, *Unsupported* e *Wrong*.

	I	II
Right	21	25
ineXact	1	3
Unsupported	1	1
Wrong	4	7
Total	27	36

Tabela 5.7: Número de respostas validadas.

Com base nestes resultados, pode-se calcular as quatro medidas apresentadas em 5.1 que visam avaliar sistemas de validação - *precision*, *recall*, medida-F e *qa_accuracy*. A Tabela 5.8 indica tais resultados obtidos através das duas métricas usadas.

	I	II
precision	$21/27 \approx 0.77$	$25/36 \approx 0.69$
recall	$21/32 \approx 0.66$	$25/32 \approx 0.78$
f-measure	≈ 0.71	≈ 0.73
qa_accuracy	$21/73 \approx 0.29$	$25/73 \approx 0.34$

Tabela 5.8: Resultados obtidos para 73 perguntas do CLEF 2008.

Como se pode observar, onze das 32 respostas correctas, utilizando a primeira métrica do sub componente de validação de fundamentação da resposta, e sete, utilizando a segunda, não foram validadas. Em quatro dessas respostas incorrectamente não validadas o sub-componente de validação do tipo da resposta retornou o valor 0 por não haver nenhuma semelhança com o tipo da resposta submetida a validação e o tipo esperado. Tal resultado, tem a capacidade de anular o obtido no outro sub componente e faz com que a resposta não seja validada. A razão pela qual o sub componente de validação do tipo da resposta retorna o valor 0 para estas respostas, está relacionada com o facto de o XIP não etiquetar tais respostas com o tipo ou sub-tipo esperado com base na análise da pergunta. Por exem-

plo, na pergunta *"Quantas províncias tem a Ucrânia?"*, a resposta retornada e considerada correcta no fórum de avaliação CLEF 2008 foi "24". A etiqueta do XIP que caracteriza o tipo esperado da resposta é *QUANT*, contudo o XIP não marca a resposta com tal etiqueta nem com nenhuma outra pertencente ao sub-tipo da classe que representa os quantificadores. Uma solução possível passa por relaxar as regras de comparação entre o tipo esperado e o tipo da resposta a ser validada. Neste caso específico, poderia-se admitir que o tipo *NUM* (número), associado pelo XIP à palavra "24", tenha uma relação com o tipo *QUANT* (quantidade), evitando que o componente de validação do tipo da resposta retornasse 0. Contudo, diminuir a exigência no emparelhamento entre os tipos da resposta candidata e do esperado aumentará certamente o número de validações de respostas erradas.

Comparando as duas métricas usadas no componente de verificação da fundamentação da resposta, pode-se concluir que a segunda é mais branda com base em dois factos:

- não há nenhuma resposta que seja validada com base na primeira métrica e rejeitada pela segunda;
- através da segunda métrica são validadas mais nove perguntas do total das 73 do conjunto de teste considerado.

Tal situação é justificada pelo facto de que a primeira métrica castiga com maior impacto os textos de suporte que não contenham determinadas entidades mencionadas ou vocábulos auxiliares presentes na pergunta. Por sua vez, a segunda métrica beneficia textos de suporte que contenham entidades mencionadas próximos da resposta candidata, deixando para segundo plano a possibilidade de outros vocábulos da pergunta não se encontrarem contidos no referido texto de suporte.

6 Conclusão

O trabalho realizado tem como objectivo primordial tornar o sistema QA@L^{2}F mais eficaz na busca de possíveis respostas candidatas e selecção das mesmas. Para a concretização de tal objectivo decidiu-se fazer um levantamento dos problemas mais graves do sistema, através de uma análise do mesmo, a partir da versão datada de Outubro de 2007. Fez-se também um estudo a vários sistema de pergunta-resposta e de validação de resposta, descrito em 3, que utilizam técnicas capazes de dar resposta aos problemas acima referidos e descritos em 2.

Na avaliação feita ao sistema, detalhada em 5, pode-se concluir que houve uma melhoria nos resultados de 14% para 20.5%. A melhoria da Técnica de Extracção de Força Bruta, explicada em 4.2, e a implementação da nova Técnica de Extracção Baseada em Distâncias, em 4.3, foram os grandes responsáveis destes resultados, abrangendo 60% das perguntas que o sistema respondeu correctamente. Apesar de ser um aspecto positivo, tais números provam que o QA@L2F necessita de se desenvolver ainda mais e que sistemas de *question-answering*, e outros na área de Língua Natural, exigem um trabalho contínuo ao longo do tempo para que se possam abordar os mais variados desafios que esta área propõe tais como a variedade de formulações linguísticas para construir a mesma pergunta e a ambiguidade linguística.

O estudo feito permitiu também a implementação de um módulo de validação de resposta e integração com o sistema QA@L^{2}F. O referido módulo descrito em 4.4 mostra-se útil para evitar o retorno de respostas do tipo diferente do esperado bem como avaliar a qualidade de cada resposta candidata o que possibilita uma escolha mais fundamentada da resposta final que o sistema retorna. Contudo, conclui-se que para sistemas de *question-answering*, ainda numa fase inicial, deve-se dar prioridade ao desenvolvimento em profundidade de um conjunto restrito de estratégias de extracção de resposta de forma a aumentar a eficácia do sistema, alargando o leque de perguntas às quais o sistema retorna uma resposta, isto é, aumentando a medida de cobertura do sistema. Quanto maior for a taxa de cobertura de um sistema de *question-answering*, maior é a utilidade de um módulo de validação de resposta a integrar nesse sistema.

Para trabalho futuro recomenda-se que se adopte uma de duas abordagens possíveis:

- Desenvolver com maior profundidade estratégias já implementadas.
- Implementar novos módulos de extracção de resposta.

No primeiro caso, um dos possíveis caminhos a ter em conta, é a melhoria da produção de dependências concebidas a partir da cadeia de Processamento de Lingua Natural. Como é explicado na secção 4.2, este trabalho visou a produção de mais dependências, que relacionassem respostas candidatas a conceitos-chave de perguntas, criadas a partir de regras com um grau de complexidade maior. No entanto, será sempre possível dotar o sistema de novos padrões para captar outras dependências ainda mais complexas. Pode-se, também a este nível, simular capacidade de raciocínio no sistema, produzindo, para esse efeito, dependências onde, por exemplo, se relacione graus de parentesco. Isto é, fazer com que o sistema, ao se produzir a dependência que contém a informação de que X é pai de Y, criasse automaticamente outra com a informação de que Y é filho de X.

Optando-se pela segunda abordagem acima referida, uma ideia que pode ser interessante passaria por alargar a base de conhecimento, isto é o corpus jornalístico disponível e os artigos da Wikipédia, a outros recursos, nomeadamente a Internet. Esta possível estratégia tinha, contudo, o inconveniente de diminuir significativamente a fidedignidade da informação contida nesta nova fonte que o sistema usaria.

Bibliography

Ait-Mokhtar, S., Chanod, J.-P., & Roux, C. (2001, October). A multi-input dependency parser. In *Proceedings of the seventh iwpt (international workshop on parsing technologies).* Beijing, China.

Amaral, C., Fugueira, H., Mendes, A., Mendes, P., & Pinto, C. (2005). A workbench for developing natural language processing tools.

Bouma, G., Kloosterman, G., Mur, J., Noord, G. van, Plas, L. van der, & Tiedemann, J. (2007). Question answering with Joost at clef 2007. *Working Notes for the CLEF 2007 Workshop*.

Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., et al. (2006). Priberam's question answering system in a cross-language environment. *Working Notes for the CLEF 2006 Workshop*.

Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., et al. (2007). Priberam's question answering system in a cross-language environment. *Working Notes for the CLEF 2007 Workshop*.

Castillo, J. J. (2008). The contribution of famaf at qa@clef2008 answer validation exercise. *Working Notes for CLEF 2008 Workshop*.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Ferrández Óscar, Micol, D., Muñoz, R., & Palomar, M. (2007). The contribution of the university of alicante to ave 2007. *Working Notes for CLEF 2007 Workshop*.

Garcia-Cumbreras, M. A., Perea-Ortega, J. M., Santiago, F. M., & Ureña-López, L. A. (2007). Sinai at qa@clef 2007. answer validation exercise. *Working Notes for CLEF 2007 Workshop*.

Glõckner, I. (2007). University of hagen at clef 2007: Answer validation exercise. *Working Notes for CLEF 2007 Workshop*.

Iftene, A., & Balahur-Dobrescu, A. (2007). Answer validation on english and romanian languages. *Working Notes for CLEF 2007 Workshop*.

Iftene, A., & Balahur-Dobrescu, A. (2008). Answer validation on english and romanian languages. *Working Notes for CLEF 2008 Workshop*.

Jacquin, C., Monceaux, L., & Desmontils, E. (2008). The answer validation system prodicosav. *Working Notes for CLEF 2008 Workshop*.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*.

Lin, D. (1998). An information-theoretic definition of similarity. *Procedings of the 15th International Conference on Machine Learning*.

Mendes, A. (2007). *L2f@qa: primeiros passos*. Unpublished master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal. (work in progress)

Monceaux, L., Jacquin, C., & Desmontils, E. (2005). The query answering system prodicos. *Working Notes for CLEF 2005 Workshop*.

Moriceau, V., Tannier, X., Grappy, A., & Grau, B. (2008). Justification of answers by verification of dependency relations - the french ave task. *Working Notes for CLEF 2008 Workshop*.

Pardal, J. P., & Mamede, N. J. (2004, November). *Terms Spotting with Linguistics and Statistics.*

Ribeiro, R., Mamede, N. J., & Trancoso, I. (2003). Using Morphossyntactic Information in TTS Systems: comparing strategies for European Portuguese. In *Computational processing of the portuguese language: 6th international workshop, propor 2003, faro, portugal, june 26-27, 2003. proceedings* (Vol. 2721). Springer.

Rodrigo Álvaro, Peñas, A., & Verdejo, F. (2007). Uned at answer validation exercise 2007. *Working Notes for CLEF 2007 Workshop*.

Rodrigo Álvaro, Peñas, A., & Verdejo, F. (2008). The efect of entity recognition in the answer validation. *Working Notes for CLEF 2008 Workshop*.

Sacaleanu, B., Neumann, G., & Spurk, C. (2007). Dfki-lt at qa@clef 2007. *Working Notes for the CLEF 2007 Workshop*.

Téllez-Valero, A., & Luis Villaseñor-Pineda, M. M. and. (2007). Inoe at ave 2007: Experiments in spanish answer validation. *Working Notes for CLEF 2007 Workshop*.

Apêndice I

A Resultados da Avaliação Clef 2008

Neste apêndice é apresentado o conjunto de teste da edição de 2008 do CLEF, constituído por 200 perguntas. Cada pergunta tem associada a informação relativa ao seu *id*, ao grupo a que pertence, à sua categoria - *Factoid* (F), *Definition* (D) ou *List*(L) e se tem ou não restrições temporais. Perguntas que contêm figuras de estilo como anáfora ou elipse, caracterizam-se por pertencerem ao mesmo grupo que a anterior.

ID	Grupo	Cat.	R.Temp.	Pergunta	I	II
0001	2600	F		Que animal é o Cocas?	W	W
0002	2601	F		Quem foi o criador de Tintin?	U	U
0003	2601	F		Quando é que ele foi criado?	R	R
0004	2601	F		Como se chama o cão dele?	W	W
0005	2601	F		De que raça é o cão?	W	W
0006	2602	L	✓	Diga uma escola de samba fundada nos anos 40.	W	W
0007	2603	F		Em que ano houve um terramoto no Irão?	W	W
0008	2604	F		Quanto pesa um beija-flor?	W	W
0009	2605	F		Onde ficava a Gália Cisalpina?	W	W
0010	2606	F		Quantas províncias tem a Catalunha?	W	W
0011	2607	F		Qual é a montanha mais alta do México?	W	W
0012	2607	F		E do Japão?	W	W
0013	2608	F		Onde fica Saint-Exupéry?	W	W
0014	2609	F		Qual a altura do Kebnekaise?	W	W
0015	2610	F		Quem escreveu Fernão Capelo Gaivota?	W	W
0016	2611	D		O que é um menir?	R	R
0017	2612	F		Em que ano é que Ernie Els venceu o Dubai Open?	W	W
0018	2613	F		Quantos ossos têm a face?	W	W
0019	2614	F		Quando começou o Neolítico?	W	W
0020	2615	F		Quando nasceu Thomas Mann?	R	R
0021	2615	F		E quando morreu?	W	R
0022	2616	F		A que partido pertence Zapatero?	W	W
0023	2617	D		Quem é FHC?	W	W
0024	2618	D		Quem foi Álvaro de Campos?	R	R
0025	2618	L		Diga uma das suas obras.	W	W
0026	2619	F		Os tucanos são membros de que partido?	W	W
0027	2620	D		Quem foi Pierre Larousse?	X	R
0028	2621	D		O que é a Brabançonne?	W	W
0029	2622	F		Onde vivem as aves tucanos?	W	W

ID	Grupo	Cat.	R.Temp.	Pergunta	I	II
0030	2623	F		Em que estado brasileiro habitam os tucanos?	W	W
0031	2624	F		Quem disse "alea iacta est"?	W	W
0032	2624	F		Ao atravessar que rio?	W	W
0033	2625	F		Que político é conhecido como Iznogoud?	W	W
0034	2626	D		O que é uma cítara?	R	R
0035	2627	D		O que era o A6M Zero?	W	W
0036	2628	L		Diga um gás nobre.	W	W
0037	2628	L		E um não-metal	W	W
0038	2629	F		Qual é o asteróide número 4?	W	W
0039	2630	F		Qual a capital da Picardia?	W	W
0040	2631	F		Quando reinou Isabel II de Castela?	W	W
0041	2632	D		Quem é Narcís Serra?	W	W
0042	2633	F		Como se chamava o cavalo do Dom Quixote?	W	W
0043	2634	F		Qual é a capital do estado de Nova York?	W	W
0044	2635	L		Quais são as províncias da Irlanda?	X	X
0045	2636	F		Que instrumento tocava Ringo Starr?	W	W
0046	2637	F		Que papa sucedeu a Leão X?	W	W
0047	2638	F		Quem é o pato mais rico do mundo?	W	W
0048	2639	F		Quem são os sobrinhos do Pato Donald?	W	W
0049	2639	F		E a namorada dele?	W	W
0050	2639	F		Qual a profissão dele?	W	W
0051	2640	D		O que é a paella?	R	R
0052	2641	L		Que países abrange a Lapónia?	W	W
0053	2642	D		O que é a açorda?	X	R
0054	2643	D		O que é o feta?	R	R
0055	2643	F		De que país é originário?	W	W
0056	2644	F		Em que ano foi construída a sinagoga de Curação?	W	W
0057	2645	F		Com que idade o Mequinho foi campeão brasileiro de xadrez?	W	W
0058	2646	F		Quem dirigiu o Japão durante a Segunda Guerra Mundial?	W	W
0059	2647	F		Quantas repúblicas formavam a URSS?	W	W
0060	2648	F		Em que país fica a Ossétia do Norte?	W	W
0061	2648	F		E a Ossétia do Sul?	W	W
0062	2649	F		Qual a largura do Canal da Mancha no seu ponto mais estreito?	W	W
0063	2650	F		Quem criou Descobridores de Catan?	W	W
0064	2651	F		Quem é o santo patrono dos cervejeiros?	W	W
0065	2651	F		E do pão?	W	W
0066	2652	F		O que é o jagertee?	W	W
0067	2653	F		Qual a envergadura de um milhafre-preto?	W	W
0068	2653	F		Quanto é que ele pesa?	R	R
0069	2653	F		Que tipo de ave é?	W	W
0070	2654	F		Quantas províncias tem a Ucrânia?	R	R
0071	2655	F		Que partido foi fundado por Amílcar Cabral?	U	U
0072	2656	F		Quantos filhos teve a rainha Cristina da Suécia?	W	W
0073	2657	F		Quem é o dono do Chelsea?	W	W
0074	2658	F	✓	Quantos habitantes tinha Berlim em 1850?	W	W
0075	2658	F	✓	Quantos tem hoje em dia?	W	W
0076	2659	D		O que é o ICCROM?	X	X

ID	Grupo	Cat.	R.Temp.	Pergunta	I	II
0077	2659	F	✓	Quantos estados membros tinha em 1995?	R	R
0078	2659	F		Onde tem a sua sede?	W	W
0079	2660	F		Quantas vezes ganhou Portugal a Taça Davis?	R	R
0080	2661	D		O que é o IPM em Portugal?	W	W
0081	2662	F		Quem foi o último rei de Portugal?	W	W
0082	2662	F		Em que período foi ele rei?	W	W
0083	2662	F		Em que barco ele embarcou para o exílio?	W	W
0084	2663	L		Diga uma batalha ocorrida durante a Guerra dos Cem Anos	W	W
0085	2664	F	✓	Quantos votos teve o Lula nas eleições presidenciais de 2002?	W	W
0086	2664	F		Quando é que ele tomou posse?	W	W
0087	2665	F		Quem era o pai de Carlomano?	W	W
0088	2666	D		Quem foi Baden Powell de Aquino?	W	W
0089	2667	F		Quem escreveu o Livro da Selva?	W	W
0090	2667	F		Quem é a personagem principal do livro?	W	W
0091	2668	F		Em que ilha fica Sapporo?	R	R
0092	2669	F		Quem fundou a escola estóica?	W	W
0093	2670	F		Quais são as regiões da Bélgica?	W	W
0094	2671	F		Qual é o 31º estado dos Estados Unidos?	W	W
0095	2671	F		E o 37º?	W	W
0096	2672	D		O que era a RSFSR?	W	W
0097	2673	F	✓	Quantos atletas participaram nos Jogos Olímpicos de 1976?	W	W
0098	2673	F	✓	Em que país se realizaram?	W	W
0099	2673	F	✓	E em que cidade?	W	W
0100	2674	D		O que é um berimbau?	R	R
0101	2675	L		Que países fazem fronteira com a Itália?	W	W
0102	2676	F		Como se chama o xadrez japonês?	W	W
0103	2677	F		Qual é a temperatura do zero absoluto?	X	X
0104	2678	F		Quem era a deusa da sabedoria?	W	W
0105	2679	F		Que rio banha Paris?	R	R
0106	2680	F		Qual o comprimento do Spree?	U	X
0107	2681	F		Qual é a capital do Cazaquistão?	U	U
0108	2681	F		E a sua maior cidade?	W	W
0109	2682	F		Quem é o actual presidente da Guatemala?	W	W
0110	2682	F	✓	Qual era o cargo dele em 1991?	W	W
0111	2683	F		Quantas faixas tem a bandeira dos Estados Unidos?	R	R
0112	2684	L		Quais as cores da bandeira da Hungria?	W	W
0113	2685	F		Quando ocorreu a batalha de Torres Vedras?	W	W
0114	2686	F		Quem é o papa dos Infiéis?	R	R
0115	2687	D		O que é VRML?	R	R
0116	2688	F		Onde está a Arca da Aliança?	W	W
0117	2689	F		Como se chamava o Huambo durante a era colonial?	W	W
0118	2690	F		Qual é a língua oficial do Egipto?	U	U
0119	2691	L		Quais os submarinos da Marinha Brasileira?	W	W
0120	2692	F		Em que guerra combateu Joana de Arc?	W	W
0121	2692	F		Onde é que ela foi queimada?	W	W
0122	2692	F		Quando?	W	W
0123	2692	F		Que idade tinha ela?	W	W

ID	Grupo	Cat.	R.Temp.	Pergunta	I	II
0124	2693	F		Desde quando está Fidel Castro no poder?	W	W
0125	2693	F		Quando é que ele nasceu?	W	W
0126	2693	F		Quem é o irmão dele?	W	W
0127	2694	D		O que são os forcados?	W	W
0128	2695	F		Quando foi assinado o Tratado de Zamora?	W	W
0129	2696	D		O que é o fogo de São Telmo?	R	R
0130	2697	D		O que é que é um brigadeiro?	R	R
0131	2698	F		Quem inventou o forno de micro-ondas?	W	W
0132	2699	F		Qual a nacionalidade de Nicole Kidman?	W	W
0133	2700	F		Quem patenteou o primeiro telégrafo sem fios?	W	W
0134	2701	F		Qual é a companhia francesa de caminhos-de-ferro?	W	W
0135	2702	D		O que é a Feplam?	R	R
0136	2703	F		Qual a dotação do Prémio Cervantes?	W	W
0137	2703	F	✓	Quem é que ganhou o prémio em 1994?	W	W
0138	2704	F		Quem são os co-príncipes de Andorra?	W	W
0139	2705	F		Que tipo de tecido é o damasco?	W	W
0140	2706	F		Quantos jogadores tem uma equipa de voleibol?	W	W
0141	2707	F		Quando é que viveu Zenão de Eleia?	W	W
0142	2708	F		Qual é a área da Groenlândia?	W	W
0143	2709	F		Quem foi a primeira mulher no espaço?	W	W
0144	2709	F		E a segunda?	W	W
0145	2710	L		Diga um jornal libanês.	W	W
0146	2711	F		Quantos refugiados haitianos estão na base de Guantanamo?	W	W
0147	2712	F		Quando foi fundado o Vasco da Gama?	W	W
0148	2712	F		Por quem foi fundado?	W	W
0149	2713	F		Quando nasceu Vasco da Gama?	W	W
0150	2713	F		Onde é que ele morreu?	W	W
0151	2714	F		Em que distrito fica Sines?	R	R
0152	2715	F		Qual é a capital de Dublin?	W	W
0153	2716	F		Em que ano é que Halle Berry venceu o Óscar?	W	W
0154	2717	F		Por que estados corre o Havel?	X	X
0155	2718	L		Diga um escritor irlandês.	R	R
0156	2719	D		Quem foi Carl Barks?	W	W
0157	2719	F		Onde é que ele nasceu?	R	R
0158	2719	F		Quem eram os pais dele?	W	W
0159	2720	D		O que é um kilt?	R	R
0160	2721	F		Quem realizou ≪Os Pássaros≫?	W	W
0161	2722	F		Quantos filmes realizou Jean Vigo?	W	W
0162	2722	L		Diga um desses filmes.	W	W
0163	2723	F		Qual o comprimento da Ponte do Øresund?	W	W
0164	2724	F		Que companhia está baseada no Aeroporto Ben Gurion?	W	W
0165	2725	F		Que navio americano foi afundado em Pearl Harbor	W	W
0166	2725	F		E que navio japonês?	R	R
0167	2726	D		O que é o Crescente Fértil?	R	R
0168	2727	L		Diga um clube de futebol de Campinas.	W	W
0169	2727	L		E um de Belo Horizonte.	W	W
0170	2728	F		Qual a capital do Mato Grosso?	U	U
0171	2729	F		Quem foi o oitavo marido de Elizabeth Taylor?	W	W

ID	Grupo	Cat.	R.Temp.	Pergunta	I	II
0172	2729	F		Quando é que eles se casaram?	W	W
0173	2729	F		Qual é a nacionalidade dela?	W	W
0174	2730	F		Quantos gêneros tem o alemão?	W	W
0175	2730	F		E quantos tem o romanche?	W	W
0176	2731	F		Quanto tempo reinou Ramsés II?	W	W
0177	2731	F		Quando começou o seu reinado?	W	W
0178	2731	F		Ele ordenou a construção de que templos?	W	W
0179	2732	F	✓	Que se passou a 9 de Novembro de 1991?	W	W
0180	2733	F		Quantos actos tem a ópera Verdi da Aida?	W	W
0181	2733	F		Quem escreveu o libretto dessa ópera?	W	W
0182	2733	F		Quando é que estreou a ópera?	W	W
0183	2734	F	✓	Quem se tornou lider do Partido Quebequense em 2005?	W	W
0184	2735	F		Qual é a maior cidade do Canadá?	U	U
0185	2736	D		O que é o Gil Vicente FC?	W	W
0186	2737	D		Quem foi Gil Vicente?	R	R
0187	2738	F		Quem foi o ”pai do teatro português”?	W	W
0188	2739	F		Qual a área do Parque Estadual Guariba?	W	W
0189	2739	F		Quando foi criado o parque?	W	W
0190	2740	D		O que é a Torre do Tombo?	R	R
0191	2740	F		Onde fica?	R	R
0192	2741	F		Que país faz fronteira com Cuba?	R	R
0193	2742	F		Qual é o comprimento do metro de Coimbra?	R	R
0194	2743	F		Quantas esposas tinha Ngungunhane?	W	W
0195	2743	F		Como é que se chamava o filho dele?	W	W
0196	2744	F		Qual é a capital de Cuba?	R	R
0197	2745	F		Quem criou o primeiro alfabeto?	W	W
0198	2746	F		Quando é que Porto Rico se tornou um estados dos EUA?	R	R
0199	2747	F		Onde fica Livorno?	R	R
0200	2748	D		O que são os iaques?	R	R