INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Clefomania
QA@L^{2}F: Primeiros Passos

Ana Cristina Bastos Mendes

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

Júri

Presidente:	Doutor Ernesto José Marques Morgado
Orientador:	Doutora Maria Luísa Torres Ribeiro Marques da Silva Coheur
Co-orientador:	Doutor Nuno João Neves Mamede
Vogal:	Doutor Paulo Miguel Torres Duarte Quaresma

Setembro 2007

Agradecimentos

À Profª Luísa Coheur e ao Prof. Nuno Mamede pelo constante apoio e orientação durante a realização deste trabalho e escrita desta tese. E ainda por me fazerem levantar *cedíssimo* todas as terças-feiras para as reuniões semanais (e que, ao fim e ao cabo, foram extremamente importantes)...

À colega Raquel, com quem partilhei os meus anos de IST e, mais recentemente, a sala 226 do L^2F e o café no café todos os dias de manhã...

A todo o pessoal do L^2F, incluindo os companheiros de sala (até mesmo aqueles que teimam em ligar o ar condicionado com temperaturas *baixíssimas* de cerca de 20º)...

Aos meus amigos, incluindo aqueles que nem sequer fazem a mínima ideia do tema do meu trabalho...

Aos meus pais, irmão e restante família...

E ao András, *kedvesem*, porque o tempo e o espaço são relativos...

Agradeço-lhes por me terem ajudado a chegar até aqui! ☺

Lisboa, 21 de Novembro de 2007

Ana Cristina Bastos Mendes

Aos meus pais,
ao meu irmão

I was gratified to be able to answer promptly,
and I did.
I said I didn't know.

Resumo

Nesta tese apresenta-se o QA@L^{2}F, um sistema de *question-answering* que se baseia numa arquitectura composta por três módulos com funções distintas: pré-processamento do *corpus*, análise e interpretação da pergunta e extracção da resposta final.

No primeiro módulo, o sistema faz o processamento de língua natural no *corpus*, armazenando-o em bases de dados estruturadas para o efeito.

No módulo de análise e interpretação da pergunta, o sistema recolhe a informação relevante presente na pergunta (como, por exemplo, entidades mencionadas) e encaminha-a para o módulo de extracção da resposta final.

O último módulo na cadeia de processamento do sistema tem como responsabilidade devolver a resposta certa à pergunta recebida como entrada. Tem à sua disposição um conjunto de quatro estratégias de extracção de resposta que pode utilizar em função da pergunta: 1) emparelhamento de padrões linguísticos; 2) reordenação de formulações linguísticas; 3) emparelhamento de entidades mencionadas; e, 4) *brute force* com pós-processamento de língua natural. Neste último módulo funciona, também, um mecanismo de relaxamento de restrições que permite ao sistema alternar para uma estratégia diferente e menos restritiva na procura e recolha da resposta, caso a estratégia apropriada tenha falhado na descoberta da resposta.

O sistema QA@L^{2}F foi avaliado e testado no CLEF (um fórum de avaliação para sistemas de recuperação de informação) bem como numa avaliação feita *a posteriori*, no laboratório.

Abstract

QA@L^{2}F is the question-answering system presented in this thesis. Its architecture relies on three modules with different goals: corpus pre-processing, question analysis and interpretation and final answer extraction.

The first module is responsible for the natural language processing on the corpus and storing the information on databases.

In the question analysis and interpretation module, the system recovers the question's relevant information (such as named entities) and sends it to the final answer extraction module.

The system last module is responsible for returning the correct answer to the input question. It can use one of four different strategies in order to retrieve the correct answer, which are: 1) linguistic pattern matching; 2) linguistic reordering; 3) named entities matching; and 4) brute force plus natural language processing. This module has a constraint relaxation mechanism, which allows the system to switch among strategies: if no answer is retrieved using the appropriate strategy, the system relaxes and tries to find an answer in a more flexible and less constrained way.

QA@L^{2}F was evaluated and tested on CLEF (an evaluation forum for information retrieval systems) and on an afterwards evaluation in the laboratory.

Palavras-Chave
Keywords

Palavras-Chave

Análise Linguística Profunda

Padrões Linguísticos

Entidades Mencionadas

Mecanismo de Relaxamento de Restrições

Resposta Automática a Perguntas

CLEF

Keywords

Deep Linguistic Analysis

Linguistic Patterns

Named Entities

Constraint Relaxation Mechanism

Question-Answering

CLEF

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

CLEF Cross-Language Evaluation Forum

DRS Structure for the Discourse Representation

DUC Document Understanding Conference

L^{2}F Laboratório de Sistemas de Língua Falada

M-CAST Multilingual Content Aggregation System

NER Name Entity Recognition

NIST National Institute of Standards and Technology

NTCIR NII Test Collection for IR Systems

PLN Processamento de Língua Natural

POS Part of Speech

QA Question-Answering

SUMO Suggested Upper Merged Ontology

TREC Text REtrieval Conference

TRUST Text Retrieval Using Semantic Technologies

1 Introdução

Mania s. f.: espécie de perturbação ou excitação caracterizada por aferro a uma ideia fixa
– Dicionário online da Priberam `http://www.priberam.pt/dlpo/dlpo.aspx`

1.1 Motivação

Atravessamos uma época fortemente dominada pela informação que se apresenta sob diversas formas e através de diversos meios, acessível a todos e em todo o lado. No entanto, essa informação, quando existente em grande quantidade, dispersa e mal organizada, mais não é que uma amálgama de conceitos e ideias que pouca ou nenhuma utilidade tem. Torna-se, então, fundamental a existência de sistemas inteligentes capazes de analisar essa informação, processá-la e retirar dela o conteúdo relevante.

Neste contexto surgem os sistemas de Question-Answering (QA), definidos como sistemas computacionais cujo objectivo é responder com precisão a perguntas formuladas em língua natural, orientados ao utilizador e ao que este considera como útil e importante.

Um sistema de QA pode ser visto como uma "caixa preta", como representado na figura 1.1. Tendo como entrada uma qualquer questão formulada em língua natural, deve procurar a resposta exacta num *corpus* não estruturado de domínio aberto, retornando-a como saída. Ou seja, ao contrário dos motores de busca que, recebendo à entrada um conjunto de palavras-chave, devolvem ao utilizador um conjunto de documentos ou elos para documentos onde deve procurar a resposta, os sistemas de QA dão a resposta exacta a uma pergunta.

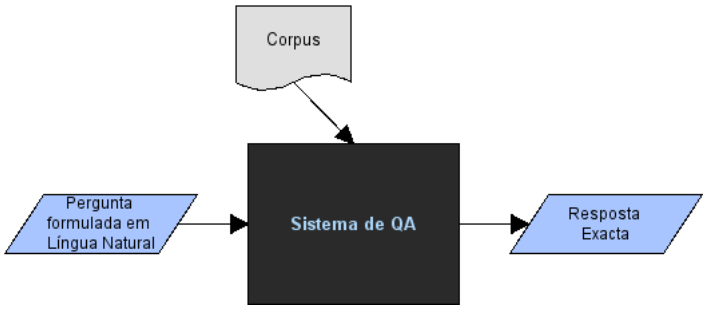

Figura 1.1: Sistema de QA como uma "caixa preta".

As linhas que se seguem ilustram uma possível interacção com um sistema de QA:

Exemplo 1:

Pergunta: Quem foi o primeiro rei de Portugal?

Resposta: D. Afonso Henriques

Pergunta: Como se chamava a sua mãe?

Resposta: D. Teresa

Pergunta: Em que ano se tornou Portugal independente?

Resposta: 1143 ■

Para a comunidade científica, os sistemas de QA são alvo de grande atenção, revelando, assim, uma importância crescente no ramo das ciências informáticas. São grandes os esforços no sentido de desenvolver técnicas que os tornem cada vez mais precisos e eficientes, respondendo cada vez mais a um maior número de perguntas. O crescente número de fóruns de avaliação de sistemas de QA (como, por exemplo, o CLEF) pode ser considerado um bom indicador da também crescente importância que estes sistemas têm tido nos últimos anos.

O projecto Clefomania nasceu com o objectivo da criação de um sistema de QA no Laboratório de Sistemas de Língua Falada (L^2F) que pudesse concorrer na edição do ano de 2007 do CLEF. Esta tese apresenta e descreve os primeiros passos dados pelo sistema QA@L^2F, desde o desenho da sua arquitectura e implementação de cada um dos seus módulos até à participação naquele fórum de avaliação de sistemas de recuperação de informação.

1.2 Problemática

A problemática da tarefa de QA prende-se, essencialmente, com o facto de estar intrinsecamente ligada à língua natural e à complexidade do seu processamento.

As próximas secções descrevem quatro dificuldades que se podem associar à tarefa de QA: a variedade das formulações linguísticas que podem ser associadas à expressão de um mesmo conceito ou ideia; a ambiguidade inerente à língua natural; a necessidade de raciocínio de forma a lidar com conhecimento não explícito; e, finalmente, a grande quantidade de informação disponível e na qual estará (ou não) a resposta correcta à pergunta.

1.2.1 Variedade de Formulações Linguísticas

Considere-se o exemplo seguinte:

Exemplo 2:

Quem inventou o saxofone?

Quem foi o inventor do saxofone?

O saxofone foi inventado por quem?

O inventor do saxofone foi quem?

A quem se atribui a invenção do saxofone? ■

Por mais ou menos intrincadas que as questões possam à primeira vista parecer, por mais ou menos informação que contenham, a conclusão final de que todas elas significam o mesmo e que se podem reduzir, por exemplo, à primeira ("Quem inventou o saxofone?") não será descabida. É fácil para um Humano este raciocínio; para um sistema computacional nem por isso. É necessário que lhe seja fornecida, por exemplo, a informação de que um inventor significa a pessoa que inventou, que um objecto é inventado por um inventor e que uma invenção é o resultado de inventar. E, tal como estas, existem inúmeras relações que devem ser dadas explicitamente ao sistema. A questão coloca-se:

Como representar a mesma informação quando descrita de formas diferentes?

1.2.2 Ambiguidade

A ambiguidade é outra das características frequentemente associadas à língua natural. De acordo com (Jurafsky & Martin, 2000), "*We say some input is ambiguous if there are multiple alternative linguistic structures that can be built from it*". [1]

Em contextos e situações do dia-a dia, não é difícil encontrarem-se exemplos em que a própria comunicação entre Humanos se revela ambígua, havendo necessidade de recorrer a uma questão do tipo "Como assim?". Muitas vezes nestes casos, por forma a eliminar a ambiguidade, o interluctor reformula o que disse, utilizando novas palavras ou uma nova estrutura sintáctica. Num sistema de QA esta interacção está muito limitada, sobretudo aquando do processamento do *corpus* (sendo, no entanto, aceitável que o sistema diga que não entendeu a pergunta do utilizador, solicitando-lhe que a reformule).

Considere-se o exemplo seguinte:

Exemplo 3:

Qual é a terceira maior cidade da Baviera? ■

[1] Das seis categorias (fonética, morfologia, sintaxe, semântica, pragmática e discurso) associadas ao estudo da língua natural descritas pelos mesmos autores, os exemplos dados focam sobretudo a ambiguidade relativa à categoria semântica.

A palavra *maior* é semanticamente ambígua. Qual é o significado que lhe é atribuído neste exemplo? Será:

- *maior*, em termos de densidade populacional?
- *maior*, em termos de superfície?
- *maior*, em termos de crescimento económico?

Naturalmente, o significado associado àquela palavra específica vai influenciar a correcção da resposta dada. Mas a ambiguidade pode ainda ser mais subtil. Veja-se o exemplo seguinte:

Exemplo 4:

Em que cidade o Mosela encontra o Reno? ■

O verbo "encontrar" está normalmente associado a um ajuntamento entre duas ou mais pessoas. Sem mais informação, a forma como a pergunta está feita pode, por exemplo, conduzir à seguinte interpretação: "Em que cidade o (personagem de filmes de cowboys) Mosela encontra o (outro personagem de filmes de cowboys) Reno (para aí travarem um duelo que há muito se tornara inevitável)?"ao invés da interpretação correcta. Neste contexto, e tendo presente que tanto Mosela como Reno são rios, a forma verbal "encontra"deve tomar o significado "desagua em": "Em que cidade o (rio) Mosela encontra o (rio) Reno?". A questão que se põe, neste caso, é:

Como interpretar correctamente o sentido dado à informação?

1.2.3 Necessidade de Raciocínio

Como, por vezes, o conhecimento e informação útil para responder a uma questão se encontra descrita implicitamente no *corpus*, é necessária a utilização de raciocínio.

O exemplo seguinte mostra uma pergunta e uma frase que contém a sua resposta. A relação entre as duas (pergunta e resposta) é estabelecida apenas com o conhecimento de que o acto de casar é bidireccional e implica o surgimento das relações "marido de" (e "mulher de") entre dois indivíduos.

Exemplo 5:

Pergunta: Com quem casou Whoppi Goldberg?

Frase 1: O marido de Whoppi Goldberg, Larry Trachtenberg,... ■

Apenas recorrendo ao raciocínio de que "se X casou com Y, então Y é marido de X", a resposta correcta (Larry Trachtenberg) é extraída.

O próximo é um exemplo em que a informação se encontra presente em duas frases, havendo necessidade de fazer a associação entre elas:

Exemplo 6:

Pergunta: Qual a antiga capital da Polónia?

Frase 1: Na história deste país, a Polónia conta com duas capitais administrativas, Varsóvia e Cracóvia,...

Frase 2: A actual capital da Polónia, Varsóvia,... ■

A resposta correcta à pergunta (Cracóvia) implica o seguinte raciocínio: "se a Polónia teve apenas duas capitais (Varsóvia e Cracóvia) e Varsóvia é a capital actual, então Cracóvia é a antiga capital da Polónia". A questão que subsiste, nestes exemplos, é:

Como lidar com o conhecimento descrito de forma não explícita?

1.2.4 Informação Disponível

O problema abordado nesta subsecção relaciona-se com a quantidade de informação que um sistema de QA tem à disposição e na qual deve pesquisar as respostas às perguntas. No caso do CLEF, são disponibilizados *corpora* onde o sistema procura a informação. Considere-se a pergunta patente no exemplo seguinte, efectuada na avaliação do ano de 2004 do fórum CLEF, bem como cada uma das suas respostas possíveis, encontradas no *corpus* disponibilizado.

Exemplo 7:

Pergunta: Quem é João Havelange?

Resposta 1: presidente da Federação Internacional de Futebol Association (FIFA)

Resposta 2: presidente da Federação Internacional de Futebol (FIFA)

Resposta 3: presidente da FIFA

Resposta 4: presidente da FIFA desde 1974

Resposta 5: único candidato assumido à sua própria sucessão na presidência da FIFA

Resposta 6: notório amigo de Castor Andrade

Resposta 7: actual presidente da FIFA

Resposta 8: líder da FIFA há já 20 anos ■

Esta pergunta tem um conjunto de 8 hipóteses de respostas correctas, mas muitas outras existem em que a resposta é apenas uma. Em todo o caso, é preciso saber localizá-la e extraí-la, independentemente do volume de informação que se tenha de analisar e processar. A questão de fundo deste problema é:

Como recuperar a informação relevante, descartando a que não interessa?

1.3 Visão Geral do Sistema

Da liberdade completa e total dada ao desenho da arquitectura do sistema, existe a ressalvar, no entanto, uma restrição conceptual: o sistema a ser desenvolvido deveria poder fazer uso de várias estratégias que lhe permitissem chegar à resposta correcta das perguntas formuladas. Fazendo uso de uma degradação gradual de requisitos (à medida que as estratégias são utilizadas sem sucesso, o sistema alarga o seu campo de acção, relaxa as suas restrições, sendo menos exigente na extracção da resposta final) o sistema tenta encontrar sucessivamente a resposta à pergunta efectuada.

Aquando do desenho da arquitectura do sistema, duas questões tiveram, também, de ser ponderadas:

- Que ferramentas de Processamento de Língua Natural (PLN) estão disponíveis no L^{2}F?
- Como fazer uso dessas ferramentas de PLN, retirando-lhes o melhor proveito e utilizando os seus resultados no sistema?

Nas duas secções que se seguem são apresentadas, respectivamente, a arquitectura subjacente ao QA@L^{2}F e as ferramentas de PLN por ele utilizadas.

1.3.1 Arquitectura do QA@L^{2}F

A figura 1.2 apresenta a arquitectura do sistema.

O QA@L^{2}F é constituído por três módulos distintos: pré-processamento do *corpus*, análise e interpretação da pergunta e, finalmente, extracção da resposta final. Os módulos referidos correspondem a cada uma das três etapas efectuadas pelo sistema de forma a devolver a resposta a uma pergunta submetida.

Em primeiro lugar, e antes da submissão da pergunta, é feito um processamento sobre o *corpus* disponível, sendo o resultado armazenado em base de dados. Este processamento faz uso da cadeia de PLN desenvolvida no L^{2}F descrita na subsecção 1.3.2. Como é visível na figura, o *corpus* pode ir directamente para a base de dados, sem recorrer a qualquer tipo de processamento.

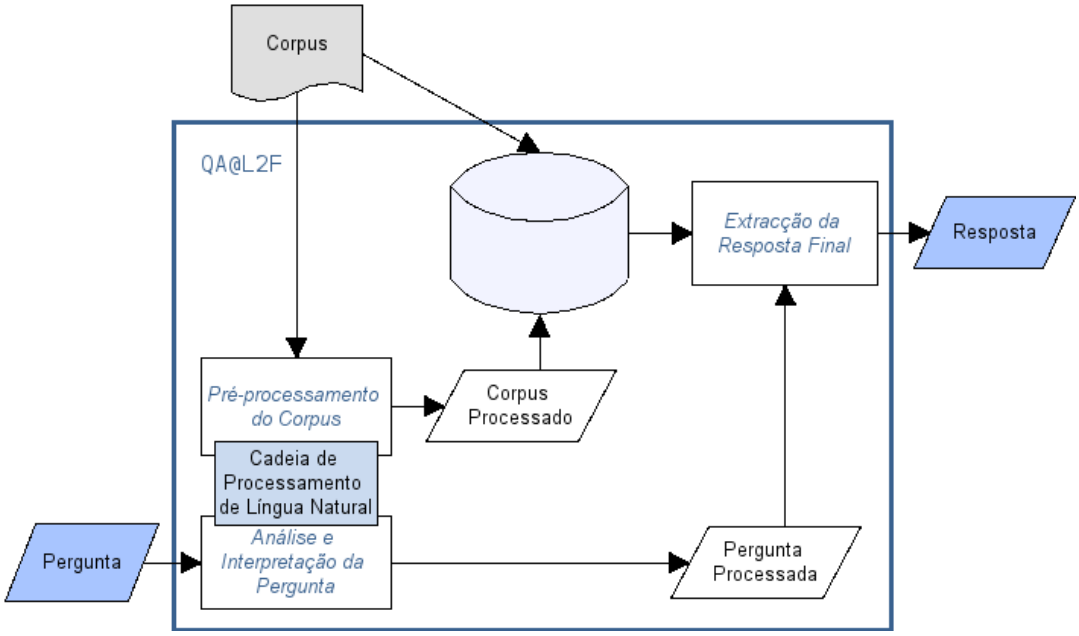

Figura 1.2: Arquitectura do QA@L^{2}F.

Sobre a cadeia de PLN assenta, também, a fase de análise e interpretação da pergunta submetida, responsável por descobrir e recolher a informação relevante presente na pergunta (tipo de pergunta, tipo de resposta esperada, entidades mencionadas) e encaminhá-la para o módulo seguinte.

A fase de extracção da resposta, última na cadeia de processamento do sistema, faz uso quer da informação relativa ao *corpus* guardada em base de dados, quer da informação recolhida acerca da pergunta, na fase anterior de análise e interpretação da pergunta. Nesta fase, o sistema tem ao seu dispor um conjunto de estratégias que pode seleccionar, mediante a pergunta submetida. Utiliza, também, um mecanismo que lhe permite adoptar estratégias diferentes para a mesma pergunta, no caso de não conseguir encontrar a sua solução. O QA@L^{2}F tem disponíveis diferentes caminhos e várias hipóteses para dar a resposta a uma pergunta.

Cada um destes três módulos será discutido em maior profundidade nos capítulos 3 e 4 deste documento.

De referir, finalmente, que a aposta durante a criação deste sistema recaíu sobre o desenho e implementação de uma arquitectura que permitisse a utilização de diferentes estratégias para responder a perguntas. Mais do que aprofundar cada um dos caminhos que o sistema tem à sua disposição, os esforços concentraram-se em analisar as perguntas passíveis de serem efectuadas ao sistema e desenvolver, a partir daí, alternativas de extracção de respostas que melhor lhes pudessem servir.

1.3.2 Cadeia de Processamento de Língua Natural

O sistema QA@L^{2}F assenta numa análise linguística profunda da pergunta e do *corpus*, efectuada pela cadeia de PLN, representada na figura 1.3.

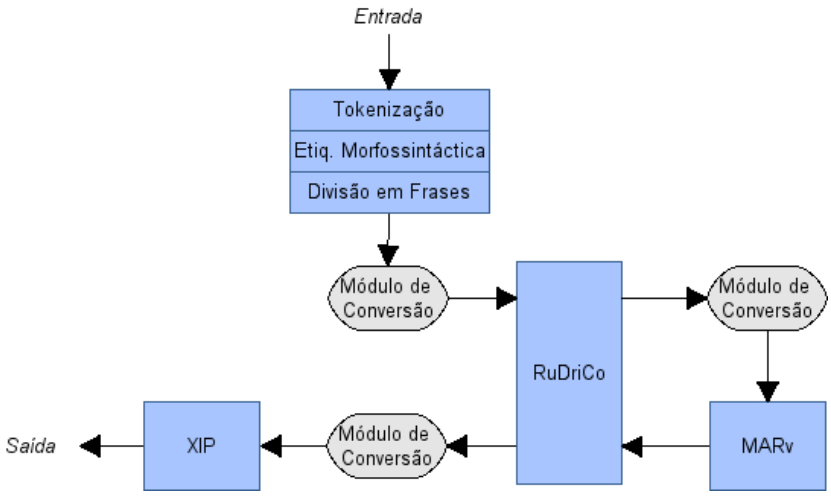

Figura 1.3: Cadeia de Processamento de Língua Natural.

A arquitectura geral da cadeia de PLN (Mamede, 2007) é definida em torno de um conjunto de ferramentas de PLN, bem como de módulos que fazem a conversão de dados de entrada/saída entre elas. As ferramentas de PLN utilizadas, e as respectivas funcionalidades, são as seguintes:

- módulo de *tokenização*, responsável por segmentar a entrada em *tokens*;
- Palavroso (Medeiros, 1995), responsável pela etiquetação morfossintáctica;
- módulo de divisão em frases, responsável pela segmentação frásica;
- RuDriCo (*Rule Driven Converter*) (Pardal & Mamede, 2004), responsável por diversas manipulações morfo-sintácticas, entre as quais, descontracção de palavras, identificação de locuções prepositivas adverbiais (na sua primeira chamada) e identificação de entidades mencionadas (na sua segunda chamada);
- MARv (Ribeiro et al., 2003; Rodrigues, 2007), responsável pela escolha da etiqueta mais provável para uma palavra;
- XIP (*Xerox Incremental Parsing*) (Aït-Mokhtar et al., 2001), responsável pelo cálculo de *chunks* e dependências.

A cadeia de PLN sofreu diversas alterações e melhorias desde a implementação e avaliação do sistema no fórum CLEF. A cadeia de PLN descrita nesta secção refere-se à existente no momento actual

de escrita da tese, tendo sido entretanto resolvidos alguns problemas de processamento, com os quais o sistema de QA teve de se debater e solucionar.

1.4 Estrutura da Tese

Esta tese encontra-se estruturada da seguinte forma: o capítulo 2 apresenta o estado da arte dos sistemas de QA, referindo também alguns dos fóruns existentes para avaliação dos sistemas; o capítulo 3 descreve as fontes de informação utilizadas como *corpus*, a forma como o sistema as organiza e armazena, e a estrutura das bases de dados construídas para o efeito; o capítulo 4 foca as fases de análise e interpretação da pergunta, e extracção da resposta final do sistema; o capítulo 5 apresenta e faz a análise dos resultados da avaliação do sistema, incluindo os provenientes da avaliação do fórum CLEF; finalmente, a conclusão deste trabalho, as suas contribuições e o trabalho futuro encontram-se no capítulo 6.

2 Estado da Arte

2.1 Introdução

Este capítulo faz referência ao estado da arte no domínio dos sistemas de QA. Inicia, na secção 2.2, com uma descrição de três fóruns de avaliação deste tipo de sistemas: TREC, NTCIR e CLEF; na secção 2.3 são abordados os sistemas de QA portugueses participantes no fórum CLEF desde 2004 e até 2006; finalmente, na secção 2.4, são apresentados sistemas de QA noutras línguas, que não a portuguesa.

2.2 Fóruns de Avaliação de Sistemas de QA

Originários de regiões distintas do globo e apesar de se encontrarem focados em idiomas também eles díspares, o objectivo destes fóruns de avaliação é semelhante: impulsionar a investigação e a discussão acerca do estado da arte e os progressos obtidos na área do processamento da língua natural, bem como promover futuras iniciativas conjuntas entre investigadores.

A avaliação dos sistemas é feito com recurso a um conjunto de exercícios distintos de avaliação, correspondendo cada um deles a uma área particular deste ramo científico, nomeadamente a de QA.

Esta secção apresenta três dos mais importantes fóruns de avaliação de sistemas de QA.

2.2.1 TREC

O Text REtrieval Conference (TREC) (WebsiteTREC, n.d.) é um projecto co-patrocinado pelo National Institute of Standards and Technology (NIST) e pelo Departamento de Defesa dos Estados Unidos da América.

As conferências do TREC são realizadas anualmente, tendo a primeira acontecido em Novembro de 1992.

Os seus objectivos são: promover a investigação na área da extracção de informação em grandes colecções de texto; facilitar a comunicação entre a indústria, o meio académico e o governo (ao proporcionar a existência de um canal de comunicação aberto para troca de ideias); acelerar a colocação no

mercado dos sistemas desenvolvidos em laboratório; e aperfeiçoar as técnicas de avaliação deste tipo de sistemas.

A avaliação no TREC consiste num conjunto de exercícios distintos, cada um focando tarefas específicas nesta área de extracção de informação. Em 2006 existiram sete tarefas, entre as quais: *SPAM Track*, que proporciona uma avaliação às actuais abordagens de filtros de e-mail; *Terabyte Track*, que consiste num teste à escalabilidade das tradicionais avaliações, baseadas em colecções de texto, com conjuntos de teste bastante maiores do que os usados presentemente.

2.2.2 NTCIR

O NII Test Collection for IR Systems (NTCIR) (WebsiteNTCIR, n.d.) consiste num conjunto de *workshops* de avaliação, cujo enfoque está em promover a investigação de tecnologias de acesso à informação, nomeadamente: extracção de informação, QA e sumarização de texto. Os *workshops* são realizados periodicamente, com intervalos de cerca de 18 meses entre si.

O projecto NTCIR teve a sua origem no continente asiático, destinando-se a avaliar sistemas de extracção de informação baseados na língua Japonesa, sendo também dada especial importância às restantes línguas do Leste Asiático, como o Chinês e o Coreano.

Os seus principais objectivos são: encorajar as tecnologias e sistemas de acesso à informação, providenciando colecções de teste de grande escala, bem como uma estrutura comum de avaliação; disponibilizar um fórum aberto a todos os grupos de investigadores interessados na comparação dos seus sistemas e na troca de resultados, opiniões e ideias; e investigar métodos e técnicas de avaliação destas tecnologias, construindo um conjunto de teste reutilizável e de cada vez maior escala, focado naquela família de línguas.

O quinto e último *workshop*, realizado em Tóquio no mês de Dezembro de 2005, considerou seis áreas para avaliação, entre as quais: *Question Answering Task*, que consiste numa simulação interactiva, na qual os sistemas devem devolver todas as respostas correctas a cada questão formulada; *Web Task*, que promove a investigação em sistemas de extracção de informação para documentos *web* de larga escala, contendo estruturas baseadas em etiquetas e *hyper-links*.

2.2.3 CLEF

O Cross-Language Evaluation Forum (CLEF) (WebsiteCLEF, n.d.) tem como principal objectivo estimular o desenvolvimento dos sistemas de extracção de informação para as línguas europeias, de forma a garantir a sua competitividade nos mercados globais. O seu ponto alto consiste num *workshop* em que todos os participantes mostram o resultado de (pelo menos) um ano de trabalho nos seus sistemas.

Desde o ano de 2000, e a partir daí uma vez por ano e por toda a Europa, estes *workshops* constituem um excelente meio para debate do estado actual da área, sendo também um forte impulsionador de novas ideias.

Em cada ano são considerados distintos exercícios para avaliação. Em 2005 e 2006 existiram as seguintes tarefas e respectivos objectivos:

1. *Mono-, Bi- and Multilingual Document Retrieval on News Collections (Ad-Hoc)*: avaliar sistemas de recuperação de documentos, recorrendo a textos provenientes de jornais;
2. *Mono- and Cross-Language Information Retrieval on Structured Scientific Data (Domain-Specific)*: testar a recuperação de informação, no domínio das ciências sociais e economia;
3. *Interactive Cross-Language Retrieval (iCLEF)*: testar sistemas orientados ao utilizador, focando-se essencialmente nos problemas de QA e recuperação de imagens em ambiente *cross-language*;
4. *Multiple Language Question Answering (QA@CLEF)*: testar sistemas de QA de domínio aberto, em ambos os ambientes monolingue e *cross-language* (uma descrição mais pormenorizada será dada mais à frente neste documento);
5. *Cross-Language Retrieval in Image Collections (ImageCLEF)*: testar a recuperação de imagens, descritas num idioma diferente daquele em que é efectuada a questão, através de técnicas de emparelhamento de texto ou de imagem;
6. *Cross-Language Speech Retrieval (CL-SR)*: avaliar sistemas de reconhecimento de discurso espontâneo falado;
7. *Multilingual Retrieval of Web Documents (WebCLEF)*: testar sistemas de recuperação de textos utilizando como único recurso a *web*, num ambiente multilingue;
8. *Cross-Language Geographical Retrieval (GeoCLEF)*: avaliar sistemas de recuperação de informação espacial, tendo em consideração o ambiente multilingue, bem como a natureza geográfica dos dados.

QA@CLEF.

A tarefa QA@CLEF tem como objectivo avaliar sistemas de QA. Os sistemas de QA recebem como entrada uma questão formulada em língua natural e devolvem a resposta precisa a essa mesma questão.

Em cada ano, o nível de complexidade associado à tarefa aumenta. O objectivo inerente a este crescendo de dificuldade prende-se com a promoção da melhoria dos sistemas, tentando fazer com que respondam a tarefas cada vez mais complicadas.

Considerando a tarefa QA@CLEF no ano de 2006, as questões pertencem a quatro categorias, dependendo da resposta esperada:

- *Factoid*: factos ou eventos (por exemplo: "Qual é a capital da França?" ou "Quem é o pai de Lisa Marie Presley?");
- *Definition*: definições de pessoas, coisas ou organizações (por exemplo: "Quem é Mário Soares?" ou "O que é a NATO?");
- *List*: listas de pessoas, objectos ou dados (por exemplo: "Quais os elementos que compõem os Beatles?" ou "Nomeie livros de José Saramago.").
- *NIL*: questões de resposta desconhecida no *corpus*. Este tipo de questão é pertinente, no sentido que os sistemas devem identificar a inexistência de resposta, ao invés de darem uma qualquer resposta errada à questão submetida.

As questões podem, também, conter restrições temporais. Estas são divididas por:

- Data (por exemplo: "Quem foi o Primeiro-ministro de Portugal em 1990?");
- Período (por exemplo: "Quantos carros foram vendidos em Espanha entre 1980 e 1995?");
- Evento (por exemplo: "Onde estudou Michael Milken antes de entrar na Universidade da Pennsylvania?").

Não sendo fornecida qualquer informação relativa ao tipo de resposta esperada (por exemplo: pessoa, organização, localização), os sistemas devem procurar as respostas num conjunto de textos constituído por artigos jornalísticos. No caso da língua Portuguesa, foram utilizados os *corpora* Público, dos anos de 1994 e 1995 (para o português Europeu), e Folha de São Paulo, também dos anos de 1994 e 1995 (para o português do Brasil).

A tarefa pode ser concretizada quer em ambiente monolingue, quer em *cross-language*. Na perspectiva monolingue, a língua na qual é formulada a questão (língua fonte) é igual à língua da colecção de textos (língua destino). Em *cross-language*, as línguas fonte e destino são diferentes.

Relativamente aos critérios de avaliação, os sistemas devem devolver a resposta precisa à questão formulada. Em 2006, foi também requisito a referência ao documento de onde a resposta foi extraída, bem como a passagem no texto que a suporta. As respostas devolvidas pelos sistemas são categorizadas segundo o seu grau de precisão:

- *Right*: se está correcta;

- *Wrong*: se está incorrecta;
- *Inexact*: se contém mais ou menos informação do que a necessária;
- *Unsupported*: se, ou não contém a identificação do documento ao qual pertence a resposta ou este está errado, ou o pedaço de texto recolhido não contém a resposta exacta.

2.3 Sistemas de QA Portugueses

Esta secção apresenta os sistemas de QA para a língua Portuguesa participantes no CLEF desde 2004 até 2006.

2.3.1 Sistema da Universidade de Évora

A Universidade de Évora esteve presente nas edições de 2004 e 2005 com o seu sistema de QA (Quaresma et al., 2004; Quaresma & Rodrigues, 2005).

O sistema desenvolvido, para ambos os anos, assenta em duas fases distintas, abordando, respectivamente, as perspectivas da extracção de informação e da recuperação de informação. A primeira fase foca a análise preliminar dos documentos; a segunda, o processamento das questões e a geração das respostas.

Cada uma destas fases traduz-se num módulo distinto, sendo cada um composto por sub-módulos.

À fase de extracção de informação correspondem os seguintes sub-módulos:

- *Análise Sintáctica*. As frases são processadas recorrendo ao analisador Palavras (Bick, 2000). Uma nova colecção de documentos, com o resultado da análise, é obtida;
- *Análise Semântica*. A nova colecção de textos é reescrita numa estrutura de representação do discurso (definida, em Kamp (Kamp & Reyle, 1993), como Structure for the Discourse Representation (DRS)) com a respectiva lista de entidades de discurso e conjunto de condições;
- *Análise Semântica e Pragmática*. Utilizando uma ontologia de domínio geral, o módulo reinterpreta a informação semântica extraída no passo anterior e cria uma nova ontologia.

Após o processamento deste primeiro módulo, o conhecimento presente no *corpus* fica representado numa ontologia de domínio específico, gerada a partir da colecção de textos, que contem os factos identificados e inferidos pelo módulo de *Análise Semântica e Pragmática*.

Por outro lado, na fase de recuperação de informação existem os sub-módulos:

- *Análise Sintáctica*. A questão é processada recorrendo ao analisador Palavras;
- *Análise Semântica*. É construída uma DRS com as respectivas entidades de discurso e condições;
- *Análise Semântica e Pragmática*. De acordo com a ontologia e com a base de conhecimento criadas, o módulo reescreve algumas condições e gera uma nova DRS.
- *Processamento de Questões*. Identifica e devolve a resposta correcta à questão submetida, através da unificação das entidades de discurso da questão com as entidades de discurso dos documentos.

O sistema de 2005 é considerado uma evolução do de 2004, sendo que a principal diferença nesses dois anos consistiu na existência, em 2005, de um módulo de pré-processamento na fase de recuperação de informação. Este permitiu resolver alguns problemas de escalabilidade do sistema, diminuindo a sua complexidade, nomeadamente nas tarefas de acesso a bases de dados relacionais usando Prolog.

Este sistema obteve, em 2004, respostas correctas na ordem dos 23,62%. O principal problema detectado neste ano foi o facto do sistema devolver respostas NIL quando a resposta existia de facto no conjunto de textos, indicando problemas quer na fase de recuperação de informação, quer relacionados com a ontologia utilizada. Em 2005, a percentagem de respostas correctas aumentou para 25%, sendo que o problema do sistema não encontrar resposta para a pergunta efectuada (quando esta de facto existia) se manteve. A origem destes problemas é descrita como estando relacionada com uma incorrecta análise pragmática.

2.3.2 RAPOSA

O sistema RAPOSA (Sarmento, 2006b) foi criado na Faculdade de Engenharia da Universidade do Porto, em conjunto com a Linguateca. Uma versão simples deste sistema foi avaliado no CLEF no ano de 2006. O protótipo considerava apenas questões do tipo *factoid*, envolvendo pessoas, locais, datas e quantidades. De um modo geral, este sistema baseia-se em técnicas superficiais de análise, bem como na anotação semântica produzida pelo sistema SIEMÊS (Sarmento, 2006c).[1]

A arquitectura do RAPOSA é constituída por quatro módulos, que funcionam de forma independente e sequencial:

1. *Analisador de Questões*. Recorre ao sistema SIEMÊS para identificar as entidades mencionadas e outros elementos de relevo presentes na questão; identifica o tipo de questão, os elementos que a constituem e o tipo de resposta pretendida; transforma a questão numa base canónica.

[1] O SIEMÊS é um sistema de reconhecimento de entidades mencionadas em português, que participou no fórum HAREM (WebsiteHAREM, n.d.), tendo obtido a melhor classificação na avaliação da abrangência dos sistemas.

2. *Extractor de Fragmentos*. Interroga a base de dados *MySQL* que contém o *corpus*, devolvendo um conjunto de fragmentos que possam conter a resposta;

3. *Gerador de Candidatos*. O sistema SIEMÊS é usado novamente nesta etapa para cada frase devolvida pelo módulo anterior, sendo assumido que a resposta será um dos elementos anotados por este sistema; este módulo devolve um conjunto de respostas candidatas, bem como as frases que as suportam;

4. *Selector de Respostas*. Selecciona uma das respostas candidatas, a frase que melhor a suporta e dá um valor de confiança à resposta escolhida. A decisão acerca de qual a melhor resposta é feita com base no número de frases que a suportam, dando-se primazia à resposta com o maior número de frases associadas.

O sistema RAPOSA foi avaliado através de duas submissões. Na primeira, o sistema foi configurado para extrair respostas candidatas usando apenas regras de emparelhamento para perguntas Quem...?, tendo obtido resultados na ordem dos 11,17% de precisão; na segunda, a estratégia foi semelhante, usando também regras mais relaxadas para perguntas dos tipos Onde...?, Quando...?, Em que ano...? e Quanto...?, na qual obteve valores de 12,77% de precisão (Peters et al., 2007).

2.3.3 GistSumm

O sistema GistSumm (Filho et al., 2006) é um sistema de sumarização desenvolvido pelo Núcleo Interinstitucional de Linguística Computacional (NILC) do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, Brasil. Participou no CLEF no ano de 2006, na tarefa de QA monolingue, utilizando a língua Portuguesa.

Baseado na premissa de que todos os textos podem ser reduzidos a uma ideia principal, descrita por uma única frase (a frase *Gist*), este sistema é constituído por três módulos independentes e que actuam no texto de forma sequencial. O primeiro é responsável pela segmentação do texto, utilizando um separador de frases automático (denominado SENTER (Pardo, 2006)). O segundo módulo tem como função pesar as frases recolhidas e, através de métodos estatísticos, escolher a frase *Gist* dentro do conjunto de todas as frases. Esta escolha pode ser realizada recorrendo a dois métodos distintos: o primeiro envolve sumarização não-orientada ao assunto (sumários genéricos), em que a frase obtida é aquela cujas palavras aparecem mais frequentemente em todo o texto; o segundo, denominado sumarização orientada ao assunto, determina a frase com o maior valor de correlação com certo assunto. Finalmente, o terceiro módulo tem por objectivo extrair outras frases de forma a serem incluídas no sumário final do texto. Estas frases são recolhidas de acordo com dois critérios puramente estatísticos: correlação com a frase *Gist* e a sua relevância para o conteúdo do texto inicial.

O sistema foi avaliado no CLEF através de duas submissões. Na primeira, o sistema devolve como resposta correcta a frase *Gist* através do método de sumarização orientada ao assunto. Na segunda submissão, é aplicado ainda um filtro à frase recolhida, de forma a torná-la mais restritiva e precisa.

Os resultados obtidos por este sistema foram de 0,00% e 1,60% de precisão (Peters et al., 2007) na primeira e segunda submissões, respectivamente. Segundo os autores do sistema, estes resultados devem-se, em grande parte, ao facto do GistSumm não ser suficiente refinado para responder correctamente a questões, os métodos estatísticos aplicados não serem os mais eficazes e os filtros aplicados restringirem o tipo de respostas a apenas quatro (quem, onde, quando e quanto), sendo que outras variações foram testadas no CLEF.

2.3.4 Priberam

A Priberam Informática participou nas edições de 2005 e 2006 do fórum CLEF com o seu sistema (Amaral et al., 2005; Cassan et al., 2006). No ano de 2005, a tarefa visada foi a QA monolingue utilizando a língua Portuguesa; em 2006, o sistema foi reformulado, passando o objectivo a focar-se nas tarefas de QA monolingue (utilizando as línguas Portuguesa e Espanhola) e *cross-language* (Português-Espanhol e Espanhol-Português).

O sistema de QA da Priberam baseia-se no módulo Português do projecto TRUST[2]. O sistema criado para o TRUST procura documentos de texto e devolve uma lista de frases pesadas contendo a resposta à pergunta formulada em língua natural.

A Priberam dispõe de um conjunto de ferramentas de processamento de língua natural, do qual fez uso no desenvolvimento do seu sistema de QA, nomeadamente:

- Recursos Lexicais:
  - Léxicos, que contêm, para cada unidade lexical: diferentes interpretações, informações referentes a Part of Speech (POS), funcionalidades semânticas, relações lexico-semânticas, domínios ontológicos e terminológicos, traduções em Inglês e Francês;
  - *Thesaurus*, que fornece um conjunto de sinónimos para cada unidade lexical;
  - Ontologia multilingue, que agrupa palavras e expressões através do seu domínio conceptual. Foi desenvolvida no contexto do projecto TRUST e constitui um meio de tradução bidireccional.
- Ferramentas de *Software*:

[2]O Text Retrieval Using Semantic Technologies (TRUST) é um projecto co-financiado pela Comissão Europeia, com o objectivo de desenvolver um motor de busca semântico multilingue capaz de processar e responder a questões formuladas em língua natural em Inglês, Francês, Italiano, Polaco e Português.

  - *SintaGest*, que permite construir e testar gramáticas independentes da língua, codificar padrões de categorização e extracção de respostas, bem como regras de desambiguação morfológica, de reconhecimento de entidades mencionadas e de produção para construir uma linguagem livre de contexto.
  - Módulos independentes para desambiguação morfológica.

- Regras Contextuais, escritas recorrendo ao *SintaGest*, que permitem o reconhecimento de entidades mencionadas e outras expressões que podem ser consideradas um único *token*;

- Outros recursos focados na tarefa de QA, nomeadamente os padrões criados no *SintaGest*. Os padrões podem ser classificados em padrões de pergunta (usados para categorizar perguntas), padrões de resposta (usados para categorizar frases na fase de indexação) e padrões de pergunta-resposta (usados para extrair uma possível resposta para uma resposta específica).

A arquitectura do sistema de QA desenvolvido pela Priberam em 2005 baseia-se na seguinte estrutura de cinco blocos:

1. *Processo de Indexação*. O conjunto de documentos é processado e indexado, em modo *offline*: para cada documento são recolhidos os domínios ontológicos e terminológicos mais importantes e, para cada frase, recolhidas as categorias de questões às quais estas possam eventualmente responder, utilizando padrões de resposta; são assinaladas palavras especiais: datas, números, entidades mencionadas, nomes próprios e expressões; cada palavra é representada pelo triplo {lema, cabeça de derivação, POS};

2. *Análise de Questões*. Recebe como entrada uma questão em língua natural, e procede à sua lematização e desambiguação morfológica; a questão é categorizada, recorrendo a padrões de pergunta, e os seus elementos *pivot* extraídos, bem como o seu lema, cabeça de derivação, POS, sinónimos e indicações se se tratar de uma palavra especial; os domínios ontológico e terminológico da questão são recolhidos;

3. *Recuperação de Documentos*. Os dados recolhidos no passo anterior são usados como entrada neste bloco: é feita uma interrogação aos documentos indexados utilizando como chave de procura o lema, cabeça de derivação e sinónimos dos elementos *pivot* da questão; cada palavra na questão e cada documento são pesados, de acordo com critérios estatísticos; o bloco devolve como saída o conjunto de 30 documentos no topo da classificação; são marcadas neste conjunto as frases contendo os *pivots* das perguntas;

4. *Recuperação de Frases*. Cada frase marcada, bem como as anteriores k e posteriores k, são analisadas e pesadas de acordo com a sua relevância semântica e proximidade com a questão; é devolvido o

conjunto de frases com classificação superior a um valor de *threshold* (estas frases são consideradas como, presumivelmente, contendo a resposta à questão inicial);

5. *Extracção de Respostas*. São aplicados padrões de pergunta-resposta a cada uma das frases, e os emparelhamentos encontrados são pesados, sendo devolvido aquele com maior classificação; as hipóteses em que o *pivot* da questão está contido na resposta são descartadas, a não ser que faça parte de uma entidade mencionada.

O esqueleto do sistema manteve-se em 2006, mas a adição da língua Espanhola implicou a importação, adaptação e reescrita dos recursos léxicos existentes (*thesaurus*, léxico, ontologia e padrões) para a nova língua, bem como das regras contextuais (quer as existentes para desambiguação morfológica, quer as de reconhecimento de entidades mencionadas). O trabalho foi facilitado em grande parte pela grande semelhança existente entre as línguas portuguesa e espanhola.

Neste segundo ano da sua participação, o sistema experimentou algumas modificações, ao nível da necessidade de resposta a questões limitadas temporalmente, da validação final das respostas extraídas e da adaptação do sistema ao ambiente *cross-language*:

- A estratégia para o problema das perguntas com restrições temporais passa pela análise das datas dos documentos e das expressões temporais existentes nos textos. No bloco *Análise de Questões*, o sistema reconhece a restrição temporal existente e converte a expressão correspondente para um formato padrão. A restrição é relaxada, passando a compreender um maior período temporal. No bloco *Recuperação de Documentos*, são feitas duas interrogações: a primeira que retorna o conjunto de documentos com datas relevantes; a segunda que retorna o conjunto de documentos contendo a expressão temporal inicial. Os 30 documentos têm de pertencer ao conjunto de documentos devolvidos pelas interrogações temporais. No bloco *Recuperação de Frases*, apenas as frases possuindo uma expressão temporal ou que pertençam a um documento devidamente datado são consideradas;

- A abordagem para melhorar a validação das respostas finais baseou-se na estratégia de apenas serem consideradas como respostas as frases contendo todos os nomes próprios e entidades mencionadas presentes na questão, devendo existir um emparelhamento de uma determinada quantidade de *pivots* nominais e verbais entre a questão e a resposta;

- Para a tarefa de QA *cross-language*, foi utilizada uma tradução directa baseada na ontologia das linguagens TRUST. São associados pesos a cada possível tradução, e escolhida a palavra que melhor classificação obtem; as restantes são consideradas como sinónimos.

Em 2005, o sistema da Priberam teve resultados de 64,5% de precisão. Apesar de ser o melhor na língua Portuguesa, alguns problemas foram apontados (sobretudo no que diz respeito à re-

sposta às questões limitadas temporalmente): as datas são indexadas de maneira diferente, consoante a forma como aparecem no texto (por exemplo, 25 de Abril de 1974 e 25/04/1974); não é tomada em consideração a data do documento. Dos constituintes do sistema, o que conduziu a um maior número de respostas erradas (33 em 71) foi o módulo *Recuperação de Documentos*.

Em 2006, o sistema conseguiu resultados de 67% de precisão na tarefa monolingue com a língua portuguesa. Na mesma tarefa, mas com a língua Espanhola, obteve 52,5%. Em ambiente *cross-language*, os resultados foram de 36% (Português-Espanhol) e 34,57% (Espanhol-Português) de precisão. O sistema desenvolvido foi o melhor em qualquer uma destas categorias de avaliação. Os principais problemas no sistema existiram ao nível da extracção das respostas candidatas (nas tarefas monolingues), quando as respostas estavam contidas em frases muito longas; na tarefas *cross-language*, os problemas verificaram-se essencialmente na fase de tradução entre as duas línguas (fonte e destino).

2.3.5 Esfinge

O sistema de QA Esfinge, desenvolvido pela Linguateca, participou nas três avaliações do CLEF que contaram com a língua portuguesa. De acordo com a Linguateca (WebsiteEsfinge, n.d.), o Esfinge é descrito como "um sistema de resposta a perguntas de domínio geral em português". O sistema baseou-se na implementação para a língua Portuguesa da arquitectura apresentada em Brill (Brill, 2003), "explorando a redundância existente na Rede onde o português é uma das linguagens mais utilizadas".

O sistema participou, em 2004, na tarefa de QA monolingue utilizando a língua Portuguesa; em 2005 e 2006 participou também na tarefa de QA multilingue, utilizando as línguas Inglesa como língua fonte e Portuguesa como língua destino.

Em 2004, a arquitectura do sistema Esfinge traduzia-se na existência dos seguintes quatro módulos (Costa, 2004):

1. *Reformulação da Questão*. São criados padrões de resposta possíveis para cada pergunta, sendo-lhes também atribuído um peso, mediante a probabilidade em encontrar uma resposta final correcta; a formatação destes padrões depende das palavras existentes na questão;

2. *Recolha de N-gramas*. Os padrões resultantes do módulo anterior são testados num repositório de informação (sendo, para tal, utilizado o motor de busca Google); as frases obtidas nesta pesquisa são extraídas e os respectivos n-gramas medidos em termos da frequência com que aparecem;

3. *Filtragem de N-gramas*. Os n-gramas são reanalisados (verificando-se a existência de características particulares, nomeadamente dígitos e palavras capitalizadas) e os seus pesos reavaliados: por exemplo, se uma questão tem como resposta uma quantidade, é dada maior importância aos n-gramas que incluem números, e descartados aqueles compostos por datas.

4. *Composição de N-gramas*. Este módulo lida com a existência de perguntas cuja solução seja composta por um conjunto de n-gramas.

A resposta final é aquela com maior peso no conjunto das respostas candidatas e que não foi descartada na fase de filtragem. Se todas as respostas foram descartadas, a resposta final é NIL.

O sistema foi submetido a duas avaliações: na primeira, a colecção de textos era apenas composta pelo conjunto de textos fornecido pelo CLEF; na segunda, o sistema utilizou também a *Internet* como fonte de informação. A percentagem de respostas correctas obtida na segunda submissão foi maior que a obtida na primeira: 15,1% contra 11,1%, respectivamente.

No ano seguinte, o sistema passou por transformações no sentido de corrigir os problemas descobertos no CLEF2004.

O módulo anterior de *Composição de N-gramas* focou-se, este ano, na tarefa de minimizar o problema do sistema devolver fragmentos das respostas correctas. Nesse sentido, o sistema, quando encontra uma possível resposta final, ao invés de parar e devolvê-la imediatamente, verifica se existe uma outra resposta que a contenha e que passe com sucesso numa nova filtragem. Se tal se verificar, esta resposta candidata torna-se a nova resposta a ser devolvida pelo sistema como final.

À arquitectura descrita anteriormente, foram adicionados, também, dois módulos (Costa, 2005):

- *Extracção de Passagens de Texto*. Este módulo recebe como entrada os padrões de resposta e submete-os ao Google, na tentativa de encontrar pedaços de texto (que possivelmente contêm a resposta) nas páginas resultantes da procura. Caso nada seja encontrado, a procura é efectuada sobre a colecção de textos do CLEF. Se mesmo assim, nenhum documento for devolvido, é feita uma derradeira tentativa de encontrar documentos utilizando critérios de procura menos específicos. Em caso de insucesso na recuperação de um documento, o sistema pára, devolvendo NIL à questão submetida;
- *Reconhecimento de Entidades Mencionadas/Classificação nos N-gramas*. Este módulo insere-se depois do módulo *Filtragem de N-gramas* descrito anteriormente. O Esfinge submete os 200 n-gramas melhor classificados ao sistema SIEMÊS e, com base na existência de tipos de resposta diferentes consoante a pergunta formulada, são novamente reordenados (se a questão é do tipo Quanto...?, os n-gramas devolvidos como quantidades pelo SIEMÊS são colocados no topo da classificação).

Para a tarefa multilingue, existe uma fase preliminar de tradução para a língua Portuguesa da questão submetida. De resto, o algoritmo para descobrir a resposta correcta não sofre quaisquer alterações.

Novamente em 2005, as duas submissões distinguiram-se apenas no facto da *Internet* ser utilizada ou não como recurso de informação. A percentagem de respostas correctas aumentou, comparativamente com o ano anterior, sendo que os resultados obtidos foram, novamente, melhores na submissão que utilizou a *Internet*: 24% contra 22%. Na tarefa multilingue os resultados foram de 13% de respostas correctas.

Em 2006, o Esfinge sofreu novas modificações, de forma a responder aos novos desafios colocados pelo CLEF: a existência de questões esperando uma lista como resposta e a necessidade de devolver, em conjunto com as respostas precisas e correctas, as frases no texto que as suportam. Neste ano, foi adicionado ao sistema e testado um novo recurso, uma base de dados de ocorrências, bem como melhorada a utilização do sistema de reconhecimento de entidades mencionadas, SIEMÊS.

Foram acrescentados, novamente, três módulos ao Esfinge (Costa, 2006):

- *Estimativa do Número de Respostas*. Este módulo nasceu da necessidade do sistema se adaptar à existência de perguntas com múltiplas respostas; usando o analisador Palavras, este módulo devolve o número de respostas pretendidas, consoante a questão submetida;
- *Base de Dados de Co-ocorrências*. O módulo encontra-se inserido depois das fases de *Reconhecimento de Entidades Mencionadas* e *Recolha de N-gramas*. Usando a base de dados de co-ocorrências BACO (Sarmento, 2006a), os pesos obtidos nas duas fases anteriores são reavaliados, de forma a tomarem em consideração a frequência dos n-gramas em *corpus* de grandes dimensões;
- *Procura por um Documento de Suporte*. Este é o último módulo do sistema; o seu objectivo é encontrar no *corpus* um documento que suporte a resposta escolhida como correcta.

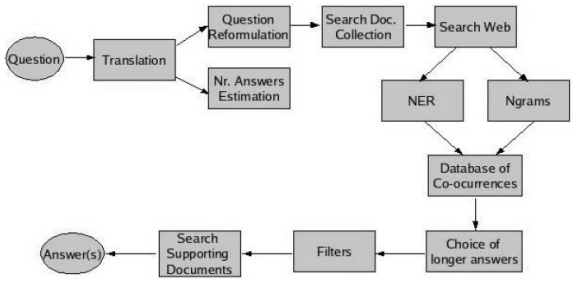

Figura 2.1: Módulos do sistema Esfinge participante no CLEF2006.

O Esfinge usa o algoritmo de QA descrito na figura 2.1.

Na tarefa monolingue, o sistema foi avaliado em duas submissões: a primeira considerou n-gramas de comprimento entre 1 e 3; a segunda considerou n-gramas com 1 a 4 palavras. Os resultados obtidos melhoraram novamente: na primeira submissão a percentagem de respostas correctas foi de 25%; na segunda, 23%. Na tarefa multiligue, os n-gramas tinham entre 1 a 3 palavras e os resultados obtidos foram de 15% de respostas correctas.

2.4 Sistemas de QA noutras Línguas

A presente secção tem como objectivo apresentar sistemas de QA para todas as restantes línguas, além da Portuguesa.

São analisados os sistemas com melhores resultados na avaliação da língua Inglesa (concorrentes ao TREC) e das línguas Espanhola, Francesa, Italiana, Alemã e Holandesa, participantes na edição do ano de 2006 do CLEF (Peters et al., 2007).

2.4.1 Inglês

A Language Computer Corporation (WebsiteLCC, n.d.) participou em 7 edições do fórum TREC, desde 1999. Desenvolveu os sistemas de QA com melhor desempenho em 6 avaliações, sendo que em 2001 o seu sistema teve os melhores resultados em duas de três tarefas (Voorhees & Buckland, 2005). Em 2006 participou, pela primeira vez, no CLEF, na tarefa de QA em ambiente *cross-language* com as línguas Inglesa, como fonte, e Espanhola, Portuguesa e Francesa, como destino. Nesta secção será apenas descrita a arquitectura do sistema desenvolvido e participante no fórum TREC no ano de 2005, que responde às tarefas monolingue Inglês-Inglês. O sistema, participante no CLEF2006, é semelhante no ambiente multilingue, contando apenas com a integração de um módulo suplementar de tradução.

O POWERANSWER-2 (Harabagiu et al., 2005), concorrente ao TREC 2005, é composto por três módulos distintos responsáveis pela análise da questão, recuperação de passagens de texto e processamento da resposta final, respectivamente. O primeiro módulo determina qual o tipo de resposta esperado para a questão submetida e extrai as palavras-chave que compõe a questão; o segundo recolhe as passagens de texto relevantes, de acordo com as palavras-chave do módulo anterior, procedendo em seguida à sua classificação e ordenação; o terceiro é responsável pela extracção da resposta final. Todos os módulos fazem uso de um analisador sintáctico, um reconhecedor de entidades mencionadas e de um sistema de resolução de referências.[3]

Do sistema fazem ainda parte dois componentes responsáveis pela eliminação de respostas erradas e prova de exactidão da resposta final. O primeiro componente tem como função procurar em documentos *web* as respostas às perguntas formuladas, utilizando padrões linguísticos e baseando-se na redundância da informação. A resposta encontrada é adicionada como palavra-chave e utilizada no módulo de extracção de respostas do *corpus*. Os resultados obtidos pela fase de extracção da resposta directamente no *corpus*, são comparados com o resultado obtido pela pesquisa na *web*, sendo dada maior

[3] A tarefa de QA no TREC apresenta algumas diferenças em relação à sua homóloga do fórum CLEF. Entre outras, os sistemas devem estar habilitados a responder a perguntas com referências entre si. Considere-se o seguinte conjunto de perguntas referente ao tópico *Jennifer Capriati*: "Quem é o seu treinador?" e "Onde é que mora?". Apesar da primeira se referir à própria Jennifer Capriati, na segunda, além desta, a entidade referida poderá ser o seu treinador.

relevância às respostas que emparelhem. Este processamento evita que respostas erradas obtidas apenas no *corpus*, apesar da sua semelhança léxica e sintáctica com a questão, sejam devolvidas pelo sistema. O segundo componente, denominado COGEX (Moldovan et al., 2003), faz uma prova por abdução[4] da exactidão da resposta. Quer a questão, quer as respostas candidatas, são transformadas, *a priori*, na sua representação em lógica, sendo utilizado conhecimento axiomático proveniente de cinco fontes: (1) uma base de dados lexical para o Inglês, a *eXtended WorNet* (WebsiteWordnet, n.d.); (2) axiomas ontológicos gerados pela ferramenta de aquisição de conhecimento JAGUAR (Bixler et al., 2005); (3) axiomas para o tratamento de padrões linguísticos (como apostos, possessivos); (4) cálculo semântico e (5) axiomas temporais disponíveis na base de conhecimento SUMO (Niles & Pease, 2001). Dependendo do valor de confiança dada a esta prova, o sistema escolhe a resposta exacta ou a resposta mais bem classificada pelo módulo de extracção de respostas (no caso da prova falhar, ou do valor de confiança ser demasiado baixo).

Tendo também de responder a questões com restrições temporais, o sistema POWERANSWER-2 segue a seguinte estratégia: detecta as datas absolutas presentes na questão e prefere as passagens de texto que emparelhem com essa restrição temporal; descobre os eventos relacionados através de sinais temporais presentes na questão e nas respostas candidatas; procede a uma unificação temporal entre a questão e as respostas candidatas e dá maior importância às respostas que emparelhem com as restrições temporais presentes na questão.

Ao contrário do que acontece no CLEF, algumas perguntas no TREC são categorizadas em tópicos: os sistemas têm, assim, a vantagem de poderem fazer uma escolha prévia da informação presente no *corpus*. Existem, no entanto, perguntas com o tópico *Other*.[5] Nestas não é dada qualquer informação extra aos sistemas. O POWERANSWER-2 apresenta três técnicas diferenciadas para fazer uma selecção de conteúdos importantes num *corpus* de grandes dimensões:

1. Utilização de padrões de questão: a selecção é feita tendo em conta as características associadas à classe do conceito em causa. Em primeiro lugar, é utilizado um classificador de Bayes para definir a classe; as suas características são-lhe depois atribuídas e realizada uma procura por essas características no *corpus*;

2. Utilização de classes de entidades: a selecção é feita com base no facto de que os pedaços de texto relevantes são os que apresentam associações com outras entidades mencionadas presentes no *corpus*. Recorre-se ao reconhecedor de entidades mencionadas para descobrir as associações entre conceitos;

[4] A abdução, introduzida por Pierce (Burch, 2007) como "*the process of forming an explanatory hypothesis*", define um facto como sendo a explicação para a existência de outro. Na fórmula `a=>b`, por dedução temos que `b` é definido como consequência de `a`; pelo contrário, na prova por abdução temos que `a` é definido como a explicação para `b`.

[5] Este tipo de perguntas é semelhante ao tipo *definition* no CLEF.

3. Utilização de padrões: a selecção é feita com recurso à pesquisa no *corpus* utilizando padrões que indicam a presença de uma definição (do conceito em causa).

O sistema obteve resultados na ordem dos 53,4% de respostas correctas, na avaliação do TREC2005.

2.4.2 Espanhol

O sistema que melhores resultados obteve no ambiente monolingue usando a língua Espanhola foi o desenvolvido pela Priberam. Este sistema está descrito na secção 2.3.4. Esta secção, prossegue com a descrição do segundo melhor sistema desenvolvido para a língua Espanhola.

O sistema (Juárez-Gonzalez et al., 2006) desenvolvido pelo Laboratório de Tecnologias da Linguagem, no México, usa estratégias de *machine learning* e *data mining* para resolver questões dos tipos *factoid* e *definition*, respectivamente.

Nas questões do tipo *factoid*, o funcionamento do sistema atravessa três fases distintas. A primeira recupera as passagens do texto com maior probabilidade de conterem a resposta. As passagens são pesadas de acordo com a semelhança entre os conjuntos de n-gramas aí contidos e existentes na questão formulada. O peso de cada passagem está relacionado com o maior n-grama pertencente à questão que aí pode ser encontrado, sendo que quanto maior o n-grama, maior o peso associado. A segunda fase é responsável por definir a classe semântica da resposta esperada para a questão efectuada. Utilizando uma abordagem baseada em expressões regulares, a ideia inerente a esta etapa do processamento consiste em reduzir o espaço de procura à classe semântica descoberta. A terceira fase, baseada numa abordagem *machine learning*, procede à extracção da resposta correcta, fazendo uso de dois módulos separados: o primeiro, que identifica todos os fragmentos do texto relacionados com a classe semântica da resposta (cada fragmento é considerado uma resposta candidata), construindo uma representação formal da resposta com base na análise do seu contexto léxico; o segundo, que selecciona a resposta com a maior probabilidade de ser a correcta, usando para tal uma abordagem *machine learning* através de um classificador de Bayes.

Na resposta a questões do tipo *definition*, o sistema é composto por três módulos. O primeiro tem como objectivo a descoberta de padrões através da pesquisa na *web* por pares (conceito, definição). A informação recolhida é normalizada e sofre a aplicação de um algoritmo de *data mining*, com o objectivo de encontrar sequências (de palavras, sinais de pontuação e outras anotações) que expressem os padrões lexicograficamente relacionados com a definição de conceitos. É aplicado, também, um filtro que escolhe os padrões mais discriminativos. O segundo módulo é responsável pela criação de um catálogo de conceitos e suas definições, resultantes da aplicação dos padrões descobertos à colecção de textos. O terceiro trata da extracção da resposta correcta para uma dada questão. O módulo funciona com a premissa de que a informação correcta se encontra em maior quantidade no catálogo do

que a informação incorrecta, e encontra-se dividido em duas sub-tarefas: a primeira, que pesquisa no catálogo todas as definições associadas ao conceito pretendido; a segunda, que escolhe a resposta correcta. Os dados são normalizados, e é aplicado novamente um algoritmo de *data mining* para obter todas as sequências máximas de palavras, sendo que a mais frequente é devolvida como resposta à questão.

Este sistema teve como resultados 51% de respostas correctas.

2.4.3 Francês

O QRISTAL (Laurent et al., 2006) é um sistema de QA *cross-language* para as línguas Francesa, Inglesa, Italiana, Portuguesa, Polaca e Checa, utilizado no projecto M-CAST.[6] O sistema foi testado no CLEF2006 nas tarefas de QA monolingue com a língua Francesa e *cross-language* (Português-Francês e Inglês-Francês).

O sistema QRISTAL para a língua Francesa utiliza extensivamente várias ferramentas de processamento de língua natural, como sejam análise sintáctica, desambiguação semântica, resolução de anáforas, detecção de metáforas, tratamento de discurso falado, extracção de entidades mencionadas e reconhecimento de conceitos e domínio. A sua arquitectura é definida em torno de dois módulos sequenciais: motor de indexação e motor de extracção de respostas.

O primeiro módulo aplica-se aos documentos da colecção de textos. Estes são particionados em blocos de dimensão 1kb e analisados sintactica e semanticamente. Desta análise resulta a criação de índices para cada: cabeça de derivação (considerado como a interpretação dada a cada palavra), nome próprio, idioma (sendo utilizado um dicionário de idiomas, com entradas como *word processing*), entidades mencionadas, classificação taxonómica, categoria, palavra-chave e tipos de pergunta (a que enventualmente o bloco possa responder) e resposta. O módulo seguinte de extracção de respostas procede à análise sintáctica e semântica da questão submetida. O tipo da questão envolvida é recolhido e cada palavra considerada *pivot* é classificada: são-lhe atribuídos diferentes pesos consoante as suas diferentes interpretações. A pesquisa pelos índices criados na fase anterior considera os pesos, sinónimos e categorias taxonómicas de cada palavra, bem como os tipos de pergunta e resposta envolvidos. Dos blocos devolvidos, os melhor classificados são, então, reanalisados: as frases aí contidas ordenadas mediante o número de palavras, entidades mencionadas e sinónimos encontrados, a presença de uma resposta que corresponda ao tipo de questão formulada e a correspondência entre categorias e domínio. O módulo procede, seguidamente, à recolha de idiomas, entidades mencionadas ou listas que emparelhem com a resposta.

[6] O Multilingual Content Aggregation System (M-CAST) (WebsiteM-Cast, n.d.) tem como objectivo o desenvolvimento de uma "infra-estrutura multilingue que permita aos produtores de conteúdos pesquisar, consultar e integrar recursos de vastas colecções multilingues de textos (e multimédia)." O sistema M-CAST baseia-se nos resultados obtidos pelo projecto TRUST.

O QRISTAL, avaliado no CLEF2006, foi o sistema que apresentou os melhores resultados na tarefa de QA monolingue, desde a edição do ano de 2003 (68,95% de respostas correctas). (Peters et al., 2007)

2.4.4 Italiano

O sistema QUASAR (Buscaldi et al., 2006) foi desenvolvido para responder à tarefa de QA monolingue com as línguas Espanhola, Francesa e Italiana. Nesta última, foi o que obteve os melhores resultados.

O sistema é composto por três módulos independentes, correspondendo cada um às tarefas de análise da questão, recuperação de passagens de texto e extracção da resposta.

O primeiro módulo, recebendo como entrada a questão, é responsável por classificá-la (informação esta mais tarde utilizada no último módulo, que aplica diferentes estratégias consoante o tipo de resposta esperada) e identificar as suas restrições. Estas restrições, recolhidas pelo anotador de POS, podem ser de dois tipos: restrição alvo, correspondendo à palavra existente na questão que poderá aparecer, no texto, mais perto da resposta correcta; restrições contextuais, que correspondem à informação que deve ser incluída na passagem de texto (recolhida posteriormente) de forma a garantir que esta contenha a resposta correcta. O segundo módulo funciona usando o sistema JIRS. O JIRS recolhe passagens de texto independentemente da língua, não utilizando qualquer conhecimento léxico e sintáctico durante o processamento. No entanto, nesta fase, o sistema não é considerado completamente independente, utilizando informação referente à classificação das perguntas e ao tipo de padrões de resposta esperada (característicos de cada língua). Este módulo recolhe conjuntos de n-gramas presentes na questão e nas passagens de texto recolhidas pelo JIRS. As passagens são pesadas de acordo com o comprimento do maior n-grama da questão aí presente. O terceiro e último módulo recebe como entrada as passagens devolvidas na fase anterior e as restrições na questão submetida. Inicia uma pesquisa nessas passagens pelo tipo de resposta esperada. As respostas candidatas são pesadas dependendo da sua posição em comparação com as restrições, sendo depois filtradas e eliminadas as que não tenham correspondência com um padrão de resposta permitido, ou que, pelo contrário, tenham correspondência com um padrão de resposta proibido. O módulo selecciona a resposta correcta através da aplicação das seguintes estratégias:

- *Voto Simples*. A resposta devolvida corresponde ao candidato mais frequente;
- *Voto Pesado*. Cada voto é multiplicado pelo peso dado ao candidato e pelo peso dado à passagem de texto que o contém;
- *Voto Máximo*. A resposta devolvida é aquela com maior peso e que ocorre na passagem melhor classificada;

- *Voto Duplo*. Semelhante ao voto simples, mas tomando em consideração os dois melhores candidatos de cada passagem;
- *TOP*. O candidato eleito pela melhor passagem de texto é devolvida.

A resposta final é a que apresenta o melhor resultado na medida de avaliação *Confident Weighted Score* (CWS)[7].

Os resultados obtidos por este sistema foram de 28,19% de respostas correctas.

2.4.5 Alemão

O QUANTICO (Sacaleanu & Neumann, 2006) é um sistema de QA de domínio aberto para as línguas Inglesa e Alemã, que faz uso de uma *framework* comum para os ambientes monolingue (Alemão-Alemão) e *cross-language* (Alemão-Inglês e Inglês-Alemão). O processo de extracção da resposta, no entanto, difere mediante o ambiente considerado e o tipo de questão submetida.

No cenário monolingue, abordado nesta secção, o sistema QUANTICO efectua uma primeira análise sintáctica (utilizando o analisador SMES (Neumann & Piskorski, 2002)) e semântica da questão, recolhendo toda a informação necessária para a fase seguinte, nomeadamente o seu tipo: *factoid* ou *definition*.[8]

O passo seguinte consiste na recuperação das passagens nos textos que possam conter as respostas. A anotação feita *a priori* do *corpus* tem aqui impacto, indicando a informação que poderá ser útil, nomeadamente entidades mencionadas e padrões linguísticos. O sistema efectua uma busca pelas entidades mencionadas e pelo repositório de estruturas sintácticas anotadas. Se a procura não for bem sucedida, é efectuada uma nova busca pela colecção de textos, desta vez procurando pelas passagens com alguma semelhança com o conceito procurado.

Na fase de extracção da resposta, o sistema pondera entre o tipo de questão recebida (*factoid* ou *definition*), efectuando uma separação da informação pretendida: estruturas simples (como entidades mencionadas, para as *factoid*) ou complexas (como frases inteiras, para as *definition*); esta fase tem como premissa o facto da redundância ser um bom indicador da aplicabilidade da informação recolhida à questão formulada.

Na última etapa do sistema, em que a resposta final é devolvida, o contexto em que se encontra a resposta é normalizado e representado num grafo; o peso associado a cada resposta é definido em termos da sua distância aos conceitos existentes no seu contexto e a distância entre estes.

[7] A medida CWS é calculada dividindo o número de estratégias que devolvem a mesma resposta pelo número total de estratégias, multiplicado por outras medidas que dependem do número de passagens devolvidas e pelo seu peso médio.

[8] O sistema QUANTICO considera apenas questões destes dois tipos, apesar de poder ser estendido para responder a perguntas do tipo *list*.

Na conferência CLEF de 2006, o sistema QUANTICO obteve valores de 42,33% de precisão na primeira submissão e de 33,33% na segunda.

2.4.6 Holandês

O sistema Joost (Bouma et al., 2006) foi desenvolvido para responder às tarefas de QA monolingue, usando a língua Holandesa, e *cross-language*, com a língua Holandesa como fonte e a Inglesa como destino.

O Joost é composto pelos componentes já habituais nos sistemas de QA: análise da questão, recuperação de passagens de texto e extracção da resposta, bem como por um módulo, denominado Qatar, que se baseia na técnica de extracção de respostas em modo *offline*. O sistema assenta, em grande parte, na análise sintáctica efectuada pelo sistema Alpino (Bouma et al., 2000). Na primeira fase do processamento, o sistema analisa a questão, com o objectivo de identificar o seu tipo e as palavras chave que a compõe. Dependendo do tipo de questão, prossegue-se para a recuperação de passagens de texto, ou é efectuada uma pesquisa pela resposta final numa tabela de factos, usando o Qatar.[9] Para as questões não respondidas pelo Qatar, o *corpus* é dividido em parágrafos. Destes, os considerados relevantes são as entradas dos módulos seguintes de extracção de respostas. A etapa final do sistema corresponde à extracção e selecção da resposta final. As respostas, quer devolvidas pelo módulo Qatar, quer pelos módulos de recuperação de informação, são ordenadas, sendo devolvida como resposta final a melhor posicionada.

As características particulares do sistema Joost são descritas em seguida:

- a componente de recuperação de informação é feita com recurso à informação linguística presente no *corpus*, nomeadamente POS, entidades mencionadas e relações de dependência;
- a extracção do *corpus* das possíveis respostas, em modo *offline*, é feita usando padrões baseados em dependências, sendo também aplicado ao *corpus* resolução de coreferências, que permite identificar entidades mencionadas relativas a conceitos presentes na frase anterior;
- na ordenação das respostas é tida em conta a semelhança sintáctica entre estas e a questão formulada, bem como a semelhança léxica entre palavras;
- as questões do tipo *definition* são respondidas usando padrões de resposta (apostos, modificadores nominais, disjunções, complementos predicativos e modificadores predicativos) bem como relações ISA;

[9] A informação presente nestas tabelas consiste num conjunto de factos reunidos *a priori*, que podem ser devolvidos como resposta final pelo sistema Qatar, se o tipo da questão corresponder a alguma categoria aí presente.

- nas questões temporalmente restritas, é atribuída uma data à frase que contém a resposta possível, de forma a evitar e eliminar conflitos entre a questão e a informação temporal presente no texto.

A avaliação do sistema Joost traduziu-se no resultado final de 31% de respostas certas.

2.5 Sumário

Ao longo deste capítulo foram apresentados diversos sistemas de QA, que testaram a sua aplicabilidade e precisão nos fóruns TREC (sistema desenvolvido para a língua Inglesa) e CLEF (todos os restantes sistemas). As abordagens seguidas resultam, maioritariamente, da combinação de técnicas de processamento de língua natural e recuperação de informação.

Voorhees (Voorhees, 2003), baseando-se no fórum TREC em 2003, sumariza o processo efectuado pelos sistemas de QA em três passos, desde a submissão da questão até à resposta final: 1) discriminação do tipo de pergunta esperada; 2) redução do *corpus* a um conjunto menor de documentos ou passagens de texto, que mais provavelmente contenham a resposta; e 3) classificação das passagens de texto que contêm a(s) resposta(s) e extracção da(s) resposta(s) a ser(em) apresentada(s) ao utilizador.

A redundância de informação existente no *corpus* é frequentemente explorada pelos sistemas. Esta característica apresenta as vantagens de: 1) aumentar a probabilidade de se encontrar a resposta que emparelhe com a questão, devido à ocorrência de múltiplas formulações linguísticas de frase semelhantes, e 2) conferir às respostas que ocorrem com maior frequência maior grau de confiança.

3 Organização da Informação

An expert knows all the answers - if you ask the right questions.

– Author Unknown

3.1 Introdução

Um sistema de QA, de domínio aberto, deve considerar, à partida, o processamento de grandes quantidades de informação, sob a forma textual, organizadas num ou em vários *corpora*. Esta premissa nasce, sobretudo, de duas características associadas ao sistema: a necessidade de dar resposta ao maior número de perguntas que lhe são colocadas e ao facto do âmbito das perguntas não estar restringido a qualquer assunto ou tema.

Tendo este conhecimento presente, várias questões surgem naturalmente:

- Qual a informação no *corpus* considerada *importante*?
- Como proceder à sua extracção?
- Onde e sob que forma guardar essa mesma informação, de maneira a ser fácil e rapidamente acedida quando tal for necessário?

Este capítulo apresenta o *corpus* utilizado pelo sistema, descrevendo o modo como é processado e guardado pelo sistema nas suas bases de dados, antes da submissão de qualquer pergunta. A secção 3.2 aborda as fontes de informação utilizadas como *corpus*, descrevendo algumas das suas características; a secção 3.3 analisa a recolha e inserção de entidades mencionadas em base de dados e a secção 3.4 refere o modo como foram recolhidos padrões linguísticos e o processo utilizado para os armazenar.

3.2 Fontes de Informação

O sistema de QA desenvolvido recorre a um conjunto de informação proveniente de diferentes fontes e, por concorrer ao fórum de avaliação CLEF, este corpus utilizado foi determinado à partida. Encontra-se dividido em textos jornalísticos e páginas provenientes de uma enciclopédia *online*: os corpora Público e

Folha de São Paulo (ambos restritos aos anos de 1994 e 1995), e a Wikipedia (na sua versão de Novembro de 2006), respectivamente.

A tabela 3.1 faz o resumo do corpus usado pelo sistema, mencionando também algumas características [1] referentes a cada fonte de informação utilizada [2].

Fonte de Informação	Público	Folha de São Paulo	Wikipedia
Tipo	Textos jornalísticos	Textos jornalísticos	Artigos enciclopédicos
Anos	1994-5	1994-5	Novembro de 1996
Idioma	Português (PT)	Português (BR)	Português (PT e BR)
Dimensão (em KB)	348.078	226.690	∼714.000
Edições	726	730	n.a.
Documentos/Páginas	106.821	103.913	312.170

Tabela 3.1: Características das fontes de informação utilizadas pelo sistema.

3.2.1 Corpus Jornalístico

O QA@L^{2}F recorre ao corpus jornalístico não processado, guardado numa base de dados independente.

A criação desta base de dados deveu-se, principalmente, às seguintes razões:

- o corpus não foi totalmente processado devido a restrições de tempo e à dimensão (564MB de texto jornalístico e as páginas da Wikipedia),
- é difícil determinar, à partida, os pedaços de informação que serão úteis na fase de extracção da resposta;
- a informação importante, porém não detectada, não se encontra presente em nenhuma base de dados estruturada;
- as ferramentas de PLN, nomeadamente as relacionadas com a classificação e identificação de entidades mencionadas, encontram-se em constante desenvolvimento. A base de dados contendo o *corpus* permite o seu pós-processamento, beneficiando do estado actual de desenvolvimento das ferramentas.

Assim, mesmo que a informação importante na resposta a uma determinada pergunta não se encontre pré-processada e disponível na tabela correspondente, ainda está acessível ao sistema. O processamento de língua natural efectuado nos textos armazenados nesta base de dados é feito depois da sub-

[1] Informação disponível em `http://acdc.linguateca.pt/aval_conjunta/CLEF/CHAVE/chave.html` e `http://ilps.science.uva.nl/WikiXML/`

[2] A dimensão da Wikipedia corresponde à dimensão das páginas XML comprimidas.

missão da pergunta e apenas se necessário, isto é, se a resposta não tiver sido encontrada na informação pré-processada.

A figura 3.1 apresenta a estrutura da base de dados que contém a informação não processada. É composta apenas por duas tabelas:

- *SNIPPET* - contém todos os parágrafos pertencentes ao *corpus* jornalístico;
- *DOCUMENT_ID* - contém os identificadores dos documentos aos quais os parágrafos pertencem.

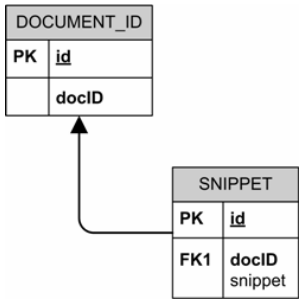

Figura 3.1: Base de dados contendo o *corpus*.

Ao *corpus* jornalístico foi ainda efectuado um processamento de língua natural, bem como o armazenamento do resultado do processamento em bases de dados, nas tabelas criadas para o efeito. As secções 3.3 e 3.4 deste capítulo abordam em maior profundidade esta etapa do QA@L^{2}F.

3.2.2 Wikipedia

As páginas provenientes desta enciclopédia *online* foram armazenados numa base de dados *MySQL* independente das utilizadas para os textos jornalísticos. Foi utilizada a versão XML da Wikipedia, denominada WikiXML collection e disponível online em `http://ilps.science.uva.nl/WikiXML/`, bem como a estrutura da base de dados disponível no mesmo *website*.

3.3 Entidades Mencionadas

O reconhecimento de entidades mencionadas (REM)[1] é definido em (Romão, 2007) como "uma subtarefa da área de extracção de informação cujo objectivo se prende com a localização e classificação de elementos átomicos num texto, tais como nomes de pessoas, organizações, locais, expressões temporais, quantidades ou valores monetários. Estes elementos contêm geralmente um nome próprio e referem-se a uma entidade específica." Assim, as entidades mencionadas são tidas como os referidos elementos atómicos capturados num texto.

[1] *Named Entity Recognition* (NER), em Inglês

3.3.1 Tipos de Entidades Mencionadas

O sistema recolhe entidades mencionadas de cada um dos tipos seguintes:

- CULTURE, relativo a obras de arte;
- CURR, relativo a unidades monetárias.
- EVENT, relativo a eventos;
- LOCATION, relativo a locais. Este tipo de entidades mencionadas pode ser associado e conter os seguintes traços:
  - `admin_area`, `continent`, `country`, `city`, `capital`, `state`, `cardinal_p`, `mountain`, `correio`, `water`, `geo_other`;
- MEASURE, relativo a quantidades. Este tipo de entidades mencionadas pode ser associado e conter os seguintes traços:
  - `length`, `area`, `volume`, `mass`, `percent`;
- ORG, relativo a organizações;
- PEOPLE, relativo a nomes de pessoas;
- PROPER, relativo a nomes próprios;
- RELATIVE, relativo a relações familiares;
- TIME, relativo a marcas temporais. Este tipo de entidades mencionadas pode ser associado e conter os seguintes traços:
  - `year`, `month`, `monthday`, `date`;
- TITLE, relativo a títulos, cargos ou profissões.

A abordagem baseada em entidades mencionadas baseia-se nos trabalhos de (Romão, 2007; Loureiro, 2007), cujo desenvolvimento se efectuou paralelamente ao do sistema QA@L^{2}F.

Esta simultaneidade teve impactos positivos e negativos no sistema de QA. Como aspecto positivo, permite que a detecção de entidades mencionadas seja feita de acordo com os requisitos do sistema. Veja-se a seguinte frase, por exemplo:

Exemplo 8:

... presidente do conselho de administração desde 2003... ■

Ao sistema de QA pode ser útil que a entidade mencionada do tipo TITLE seja apenas "presidente", ou "presidente do conselho de administração", ou "presidente do conselho de administração desde 2003" ou uma qualquer combinação destes três. A implementação paralela permite adaptar a recolha de entidades mencionadas ao sistema que as utiliza.

Contudo, a implementação parcial da recolha de entidades mencionadas, não permite ao sistema retirar o máximo proveito desta aproximação, além de poder conduzir a erros nos dados armazenados em base de dados. À altura da avaliação no fórum CLEF, durante a fase de processamento de *corpus*, apenas uma pequena parte das entidades dos tipos EVENT e ORG eram detectadas e recolhidas; por outro lado, a palavra "vitória", com letra minúscula, era categorizada como PEOPLE e, como tal, inserida indevidamente na base de dados.

3.3.2 Armazenamento em Base de Dados

Nesta abordagem, o sistema recolhe cada entidade mencionada (presente no *corpus* jornalístico), associa-a ao parágrafo onde foi encontrada e armazena-a em base de dados. Este processo é em tudo semelhante ao armazenamento de padrões linguísticos.

Considere-se, a título de exemplo, a seguinte frase de entrada (pertencente ao *corpus*)[3]:

> ***Corpus 1:***
>
> [Viaje a la Luna Luna], resposta possível a [O Cão Andaluz], de Buñuel. A viagem permite-lhe ainda um novo drama, [O Público], arrojo que nunca verá representado. De regresso a Espanha, em 1930, uma nova peça surge nos palcos de Madrid: [A Sapateira Prodigiosa].

Do processamento do ficheiro XML produzido à saída da cadeia de PLN, é gerado um ficheiro contendo, entre outras, as seguintes linhas:

Ficheiro de exemplo 1:

```
EM_START
NE_LOCATION,"Madrid "
city
capital
EM_END
EM_START
NE_LOCATION,"Espanha "
city
capital
```

[3]Recorde-se que as aspas ≪ e ≫ foram substituídas pelos parenteses [e], respectivamente

```
EM_END
EM_START
NE_NATIONALITY,"Andaluz "
EM_END
EM_START
NE_CULTURE,"[ O Cão Andaluz ] "
EM_END
EM_START
NE_TIME,"1930"
year
EM_END
```
■

Este ficheiro, identificado pelo *id* do parágrafo a que corresponde na base de dados, será processado através de um *script* na linguagem Perl e as entidades mencionadas recolhidas serão armazenadas, associadas ao parágrafo do *corpus* a que pertencem e, caso se verifique, associadas a um ou vários traços, de acordo com o tipo a que pertencem.

A figura 3.2 mostra o exemplo da estrutura da base de dados que armazena três entidades mencionadas dos tipos: LOCATION, PROPER e PEOPLE. Para as restantes entidades, a estrutura mantém-se. As tabelas *NE_<TIPO>* armazenam as entidades do tipo *<TIPO>* e as tabelas *FEATURE_NE_<TIPO>* armazenam os traços respectivos (de acordo com a lista anterior). As tabelas *SNIPPET_HAS_NE_<TIPO>* e *NE_<TIPO>_HAS_FEATURE* permitem estabelecer relações de muitos para muitos entre as tabelas que lhes estão ligadas.

A estruturação das tabelas sob esta forma de *templates* permite a criação dinâmica de interrogações à base de dados através dos *scripts*. Apenas com a indicação do tipo, a interrogação para aceder a uma entidade mencionada do tipo ORG é semelhante à interrogação para aceder a uma entidade mencionada do tipo EVENT, tendo ambas a forma:

Interrogação 1: Interrogação geral às tabelas de Entidades Mencionadas

```
SELECT *
FROM NE_<TIPO>
WHERE ne<Tipo>=?;
```
■

O mesmo acontece para o acesso a tabelas contendo os traços relativos às entidades mencionadas. A forma de recolher todas as entidades mencionadas do tipo TIME que têm o traço `year` é semelhante à forma de recolher todas as entidades mencionadas do tipo LOCATION que têm o traço `city`. Em

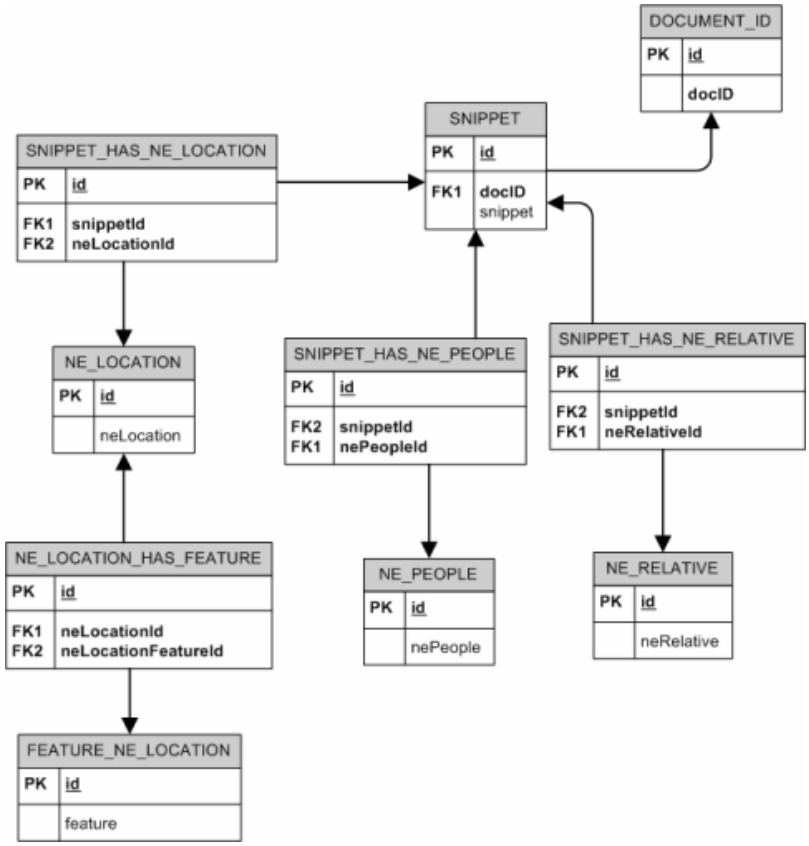

Figura 3.2: Base de dados contendo entidades mencionadas.

primeiro lugar, recolhem-se todos os *id's* das entidades mencionadas que respeitam aquela característica:

Interrogação 2: Interrogação que recolhe entidades mencionadas contendo traços específicos #1

```
SELECT ne<Tipo>Id
FROM NE_<TIPO>_HAS_FEATURE
WHERE ne<Tipo>FeatureId =
(SELECT id FROM FEATURE_NE_<TIPO> WHERE feature=<traço>)
```
■

Na posse destes identificadores, novas interrogações à base de dados podem ser efectuadas, de forma a recolher cada uma das entidades mencionadas, como pretendido:

Interrogação 3: Interrogação que recolhe entidades mencionadas contendo traços específicos #2

```
SELECT *
FROM NE_<TIPO>
WHERE id =?
```
■

Desta forma, o aumento do número de tabelas (devido ao aumento do número de entidades mencionadas reconhecidas pelo analisador), não constituirá um problema e o acesso às novas tabelas será o mesmo, sem a necessidade de alterações profundas nos *scripts*.

De referir que o objectivo desta estratégia é a fusão entre a informação presente nas questões e a informação presente no *corpus*. O sistema assume que a solução para a pergunta efectuada se encontrará nas imediações do local no *corpus* onde se encontrarem as entidades mencionadas existentes na pergunta. Assim, o *corpus* foi processado e guardado na base de dados tendo como unidade base o parágrafo (todas as entidades mencionadas fazem referência ao parágrafo onde foram detectadas). Apesar do desempenho das ferramentas de processamento de língua natural ser menor ao processar pequenos pedaços de informação (como frases) ao invés de documentos inteiros, a decisão tomada foi a de processar parágrafos, pelos seguintes motivos:

- se a janela de pesquisa de informação for muito pequena, é possível que a resposta à pergunta não esteja lá contida;
- se a janela de pesquisa de informação for muito alargada, pode haver o risco de existirem múltiplas respostas prováveis para a pergunta.

3.4 Padrões Linguísticos

Da análise do *corpus* jornalístico resultou o reconhecimento de padrões linguísticos que dão resposta directa a alguns tipos específicos de pergunta. Na génese desta pesquisa por padrões esteve, também, a análise das perguntas efectuadas nos anos de 2004 e 2006 do fórum CLEF. A ideia subjacente é a de fazer a recolha no *corpus* desses padrões, armazenar informação e, quando necessário, aceder-lhe e responder à pergunta, sem ser necessário qualquer processamento *a posteriori*.

A recolha da informação presente no *corpus*, seguindo esta abordagem, é feita inteiramente antes da submissão ao sistema da pergunta pelo utilizador. Todo o processamento é efectuado *a priori*, diminuindo o tempo de resposta à pergunta. Num sistema que tem como objectivo a interactividade com um humano, a rapidez na devolução da resposta certa é uma variável de natureza relevante. Nesse sentido, é objectivo do sistema maximizar a quantidade de relações descobertas na fase de pré-processamento do *corpus*.

A tabela 3.2 apresenta uma pergunta e um excerto retirado do *corpus* que permite a utilização desta abordagem:

Pergunta	Excerto
"Quem é Eduardo Catroga?"	*O ministro das Finanças, Eduardo Catroga, afirmou sexta-feira à noite em Coimbra...*
"Onde se situa Times Square?"	*Este «relógio da morte», instalado em Times Square (Nova Iorque)...*
"Quem realizou «Pulp Fiction»?"	*«Pulp Fiction», de Quentin Tarantino, ganha a Palma de Ouro...*
"O que é a FIL?"	*A Feira Internacional de Lisboa (FIL) abre mais uma vez...*

Tabela 3.2: Pergunta e frase contendo a resposta, permitindo a abordagem baseada em padrões linguísticos.

3.4.1 Categorias de Padrões Linguísticos

Um padrão corresponde a uma formulação linguística que captura uma relação específica existente entre dois conceitos. Os padrões foram categorizados de acordo com o tipo de relação a que dizem respeito. O sistema QA@L^{2}F possui padrões para cada uma das seguintes categorias:

Padrão da Categoria PEOPLE: Captura a relação existente entre uma pessoa e a profissão ou o cargo que ocupa. Pode, por exemplo, ser traduzido numa fórmula lógica do tipo temCargo(X, cargo) e pretende dar resposta, entre outras, ao seguinte conjunto de perguntas:

- "Quem foi X?"
- "Quem é X?"

Padrão da Categoria LOCATION: Captura a relação existente entre um conceito e o local onde se situa. Pode ser traduzido numa fórmula lógica do tipo situado(X, localização) e pretende dar resposta, entre outras, ao seguinte conjunto de perguntas:

- "Onde é X?"
- "Onde se situa X?"
- "Onde se localiza X?"
- "Onde está localizado X?"

Padrão da Categoria CULTURE: Captura a relação existente entre uma obra e o seu autor. Pode, por exemplo, ser traduzido numa fórmula lógica do tipo realizado(X, autor) e pretende dar resposta, entre outras, ao seguinte conjunto de perguntas:

- "Quem escreveu X?"

- "Quem realizou X?"
- "Quem é o autor de X?"

Padrão da Categoria STUFF: Captura a relação existente entre uma sigla ou abreviatura, com o significado dessa mesma sigla ou abreviatura. Pode ser traduzido numa fórmula lógica do tipo temSignificado(X, significado) e pretende dar resposta, entre outras, ao seguinte conjunto de perguntas:

- "O que é X?"
- "O que significa a sigla X?"
- "O que quer dizer a abreviatura X?"

3.4.2 Detecção de Padrões

Um conjunto de padrões específicos de cada categoria foram descobertos a partir da análise do *corpus*.

Na tabela 3.3 são apresentadas, para cada categoria, um padrão e o exemplo correspondente retirado do *corpus*. Nesta tabela, um padrão é uma formulação linguística composta pelos seguintes elementos: `art`, artigo; `noun[feature]`, substantivo com o traço `feature` associado; `prep`, preposição; `sigla`, sigla; `NP`, sintagma nomimal; e os restantes elementos correspondendo a sinais de pontuação.

Categoria PEOPLE	
Padrão	Exemplo

`art noun[title] noun[people],`
`art noun[title], noun[people],`
`noun[people], noun[title],`
	o ministro das Finanças Eduardo Catroga,
	o ministro das Finanças, Eduardo Catroga,
	Eduardo Catroga, ministro das Finanças,

Categoria LOCATION	
Padrão	Exemplo
`noun[location] (noun[location])` `noun[location], prep noun[location],`	Times Square (Nova Iorque) Times Square, em Nova Iorque,
Categoria CULTURE	
Padrão	Exemplo
`≪noun[culture]≫, prep noun[people]`	≪Pulp Fiction≫, de Quentin Tarantino
Categoria STUFF	
Padrão	Exemplo
`NP (sigla)` `sigla (NP)`	Feira Internacional de Lisboa (FIL) FIL (Feira Internacional de Lisboa)

Tabela 3.3: Padrões linguísticos e respectivos exemplos para cada categoria.

Os padrões deram origem a regras morfológicas, sintácticas e semânticas que emparelham directamente com as frases no *corpus*, possibilitando a extracção de relações entre conceitos.

As dependências, definidas como relações linguísticas entre duas ou mais palavras, foram criadas recorrendo ao analisador XIP. Na tabela 3.4, encontram-se, para cada categoria, as dependências criadas e um exemplo da sua aplicabilidade.

Categoria	Dependência	Exemplo
PEOPLE	(nomePessoa,cargo)	(Eduardo Catroga,ministro das Finanças)
LOCATION	(localização,localizaçãoPai)	(Times Square,Nova Iorque)
CULTURE	(obra,autor)	(Pulp Fiction,Quentin Tarantino)
STUFF	(sigla,definição)	(FIL,Feira Internacional de Lisboa)

Tabela 3.4: Dependências geradas e exemplos para cada categoria.

Para melhor se compreender a detecção de padrões linguísticos e consequente criação de dependências entre conceitos, considere-se a frase de entrada: "O ministro das Finanças, Eduardo Catroga,". A árvore sintáctica que lhe corresponde está representada na figura 3.3.

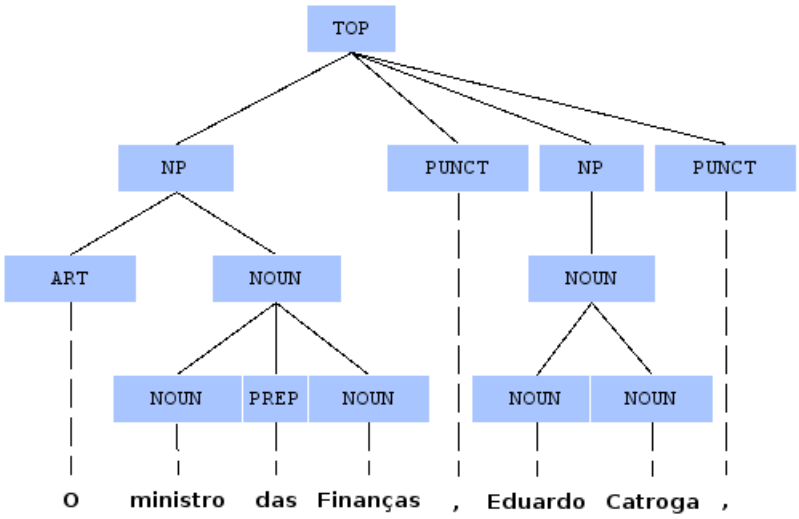

Figura 3.3: Árvore sintáctica referente à frase: "O ministro das Finanças, Eduardo Catroga,".

A seguinte regra corresponde à criação de uma dependência da categoria PEOPLE que emparelha, entre outros, com o padrão "O ministro das Finanças, Eduardo Catroga," presente na frase anterior, associando Eduardo Catroga ao seu cargo de ministro das Finanças.

Regra de Dependência 1:

```
| ?{ART*;PREP, NOUN#1[title]}, PUNCT[comma], ?{NOUN#2[people]}, PUNCT |
  PEOPLE[OK=+](#2, #1)
```

A regra é accionada se todas as seguintes condições se verificarem:

- determinado nó, independentemente da sua categoria sintáctica, é composto por um artigo (art) ou uma preposição (PREP) e um nome (NOUN), este último com o traço title associado;
- o nó seguinte é um sinal de pontuação (PUNCT) com o traço comma;
- o nó imediatamente a seguir, também pertencente a uma qualquer categoria sintáctica, é constituído por um nome com o traço people associado;
- o nó seguinte é um qualquer sinal de pontuação.

Associado a cada uma destas regras está um grau de confiança que o sistema deposita na relação, indicado na dependência gerada pelos traços OK, NORM ou KO.

As regras de dependência pertencem a um de três tipos: CERTAS, se o sistema tem grande confiança na regra (indicado pelo traço OK); NORMAIS, se o sistema tem confiança na regra (indicado pelo traço NORM); de DESESPERO, se o sistema tem uma confiança mínima na regra (indicado pelo traço KO). De referir que o tipo de dependência depende essencialmente do número de conceitos identificados. Assim, um padrão que capture a relação entre dois conceitos bem identificados pelo analisador dá origem a uma dependência certa; pelo contrário, se o padrão capturar a relação entre dois conceitos não identificados, a dependência criada é do tipo DESESPERO. Assim, a uma regra que contemple apenas conceitos bem identificados, como sejam nomes de pessoas, locais, títulos de obras culturais, o sistema dá maior confiança do que a regras que utilizem conceitos não identificados.

A seguinte regra, por exemplo, apesar de semelhante à anterior, admite que o nome entre vírgulas seja um nome próprio, sem que lhe esteja associado uma marca que o identifique como nome de pessoa.

Regra de Dependência 2:

```
| ?{ART*;PREP, NOUN#1[title]}, PUNCT[comma], ?{NOUN#2[proper]}, PUNCT |
    PEOPLE[NORM=+](#2, #1)
```

As dependências do DESESPERO verificam-se sobretudo nos padrões da categoria LOCATION, no sentido de capturarem a relação entre locais, não tendo sido nenhum deles identificado como local.

Este tipo de regras, de menor confiança, são necessárias para capturarem relações entre conceitos sem indicação explícita do seu tipo (porque não estão incluídos no léxico ou não são reconhecidos pelas gramáticas). A frase "O Presidente dos Estados Unidos, Bill Clinton," constitui um exemplo desta situação. O nome próprio Bill Clinton não é identificado e marcado com o traço people; no entanto, a regra anterior é accionada, e a relação entre a pessoa (Bill Clinton) e o seu cargo (Presidente dos Estados Unidos) é capturada nesta frase.

3.4.3 Armazenamento em Base de Dados

A passagem da informação contida nas regras de dependência para os respectivos campos na base de dados é feita de forma directa.

Do processamento da frase de entrada "o ministro das Finanças Eduardo Catroga," pelas ferramentas de PLN resulta um ficheiro XML contendo, entre outros, a seguinte dependência:

Ficheiro de exemplo 2:

```
<DEPENDENCY name="PEOPLE">
<FEATURE attribute="OK" value="+"/>
<PARAMETER ind="0" num="18" word="Eduardo Catroga"/>
<PARAMETER ind="1" num="17" word="ministro das Finanças"/>
</DEPENDENCY>
```
■

Este ficheiro será processado recorrendo a um ficheiro XSLT, de onde resultará um ficheiro de texto, identificado com o id do parágrafo a que pertence na base de dados, com a seguinte informação:

Ficheiro de exemplo 3:

```
DEPENDENCY_START
FACT_PEOPLE--99,Eduardo Catroga,ministro das Finanças.
DEPENDENCY_END
```
■

O armazenamento em base de dados é feito utilizando *scripts* na linguagem Perl, que processam ficheiros com informação semelhante à apresentada no ficheiro de exemplo anterior: introduz os componentes da relação na tabela correspondente (*FACT_PEOPLE*), faz a associação entre a relação criada e o parágrafo no *corpus* onde foi encontrado (através do nome do ficheiro), atribui à relação o seu valor de confiança (99) (dependendo do grau de confiança que o sistema tem na regra que a criou) e aumenta o contador desta relação, especificando a frequência com que é encontrada no *corpus*.

Desta forma, a tabela *FACT_PEOPLE* fica populada da forma indicada na tabela 3.5. O processamento é idêntico para as dependências das categorias CULTURE e STUFF.

A informação recolhida pela regra de dependência da categoria LOCATION, por outro lado, contém também indicação se os campos se referem a entidades mencionadas LOCATION e o *id* para estas. A adição de dois campos à tabela (*locationNeId* e *locationParentNeId*), permite a resposta a perguntas mais específicas. Além das perguntas, mais gerais, do tipo "Onde se situa X?", permite-se a resposta a

FACT_PEOPLE

id	name	title	confidence	count
1	Eduardo Catroga	ministro das Finanças	99	1

Tabela 3.5: Exemplo de entrada na tabela *FACT_PEOPLE*

perguntas do estilo "Em que cidade se situa X?". Os dois campos adicionais possibilitam a restrição das possíveis respostas, de acordo com a informação pedida. Assim, o sistema devolverá apenas como resposta as localizações que sejam cidades.

A tabela 3.6 mostra um exemplo em que tal acontece. Nova Iorque tem associado um *id* de entidade mencionada (*locationParentNeId*). Esta informação permite ao sistema descobrir que este conceito é identificado como cidade, permitindo a resposta à pergunta anterior "Em que cidade se situa X?", que de outra forma não seria dada.

FACT_LOCATION

id	location	locationNeId	locationParent	locationParentNeId	...
1	Times Square	0	Nova Iorque	10	

Tabela 3.6: Exemplo de entrada na tabela *FACT_LOCATION*

O nome da tabela cuja informação diz respeito a cada uma das quatro dependências capturadas, pertencentes a cada uma das quatro categorias descritas (tabela 3.4), tem a forma *FACT_<CATEGORIA>*. Assim, indica-se explicitamente que a tabela contém informação factual e que pode ser utilizada sem recurso a novo processamento. As tabelas intermédias *SNIPPET_HAS_FACT_<CATEGORIA>* permitem uma relação de muitos para muitos entre as tabelas *FACT_<CATEGORIA>* e *SNIPPET*. A estrutura da base de dados relativa aos padrões linguísticos está apresentada na figura 3.4. Nesta figura apresentam-se também as tabelas respectivas às entidades mencionadas LOCATION, que permitem as funcionalidades descritas no parágrafo anterior. As tabelas têm ligação aos parágrafos pertencentes aos *corpus*, e armazenados na tabela *SNIPPET*, bem como aos *id's* dos documentos a que pertencem, armazenados na tabela *DOCUMENT_ID*, já que, no contexto do fórum CLEF, é necessário que o sistema devolva o pedaço de texto do *corpus* onde encontrou a resposta certa e o *id* do documento respectivo, além da resposta à pergunta.

Antecipando algumas das perguntas que podem ser feitas ao sistema, o objectivo da criação destas tabelas é armazenar as respostas a determinadas perguntas. Evita-se, desta forma, a necessidade de se fazer qualquer tipo de processamento linguístico após a submissão da pergunta.

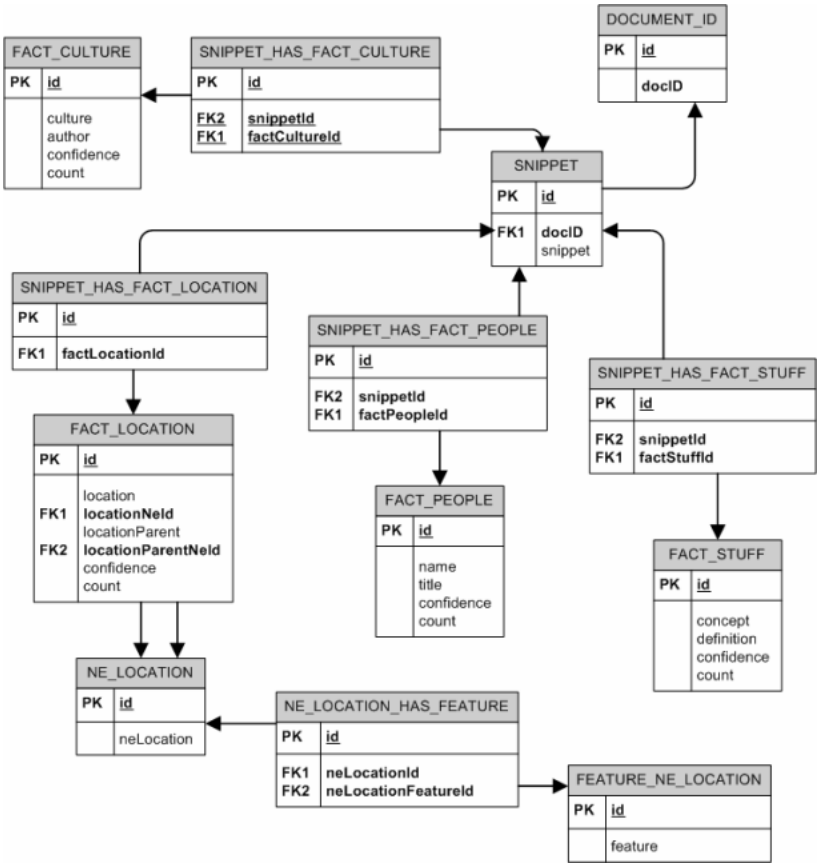

Figura 3.4: Base de dados contendo informação factual.

3.4.4 Problemática

A fase de identificação de padrões linguísticos, que permitem a recolha de relações entre conceitos, revela-se um passo importante nesta abordagem. Nesta secção abordam-se dois aspectos que é necessário ter em conta aquando da criação das regras de dependência.

Grau de Liberdade das Regras de Dependência

É preciso ter em consideração, aquando da definição das regras de dependência, qual o seu grau de liberdade. Isto é, até que ponto se devem relaxar as regras de dependência de forma a que o maior número de relações sejam apanhadas, evitando, no entanto, capturar relações inexistentes. A título de exemplo, veja-se a regra:

Regra de Dependência 3:

```
| PP{PREP[lemma:em];PREP[lemma:no];PREP[lemma:na],NOUN#1[time:~,cardinal_p:~]}, PUNCT[comma],
 {PREP[lemma:em];PREP[lemma:no];PREP[lemma:na], NP#2{NOUN[time:~]}, PUNCT |
if( ~LOCATION(#1,#2)
 & ~((#1[country:+] & #2[country:+]) |
 (#1[city:+] & #2[city:+]) |
 (#1[continent:+] & #2[continent:+])))
    LOCATION[KO=+](#1,#2)
```

A dependência LOCATION é criada entre dois conceitos se não existir uma dependência do mesmo tipo entre eles, e se ambos não contiverem os traços `country`, `city` ou `continent`. Fora estas restrições, a relação é capturada aquando da existência de uma preposição (cujo lema é "em" ou "no" ou "na"), seguida de um nome (que não pode ter sido categorizado como tempo ou como um ponto cardeal), seguido de uma vírgula, de outra preposição (com lema "em" ou "no" ou "na"), de um nome (que não pode ter sido identificado como uma marca temporal), seguido, finalmente, de um sinal de pontuação.

Apesar de ter algumas restrições, esta regra é bastante permissiva no intuito de capturar, por exemplo, locais que estejam escritos na língua Inglesa, não incluídos nos dicionários e não reconhecidos pelas gramáticas como tal. No entanto, relações inexistentes são também capturadas, como por exemplo situado(Lisboa, âmbito da Semana Europeia da Ciência), proveniente da seguinte frase presente no *corpus*.

Corpus 2:

Colóquio reúne em Lisboa, no âmbito da Semana Europeia da Ciência, especialistas europeus e norte-americanos para debaterem as relações entre a ciência e o público.

Listas de Conceitos

As listas de conceitos merecem especial relevo nesta secção. Veja-se a seguinte regra de dependência que relaciona uma pessoa e o seu respectivo cargo:

Regra de Dependência 4:

```
| ?{ART*;PREP, NOUN#2[people]}, PUNCT[comma], ?{ART*;PREP, NOUN#1[title]}, PUNCT |
    PEOPLE[OK=+](#2, #1)
```

Se a regra for aplicada em duas frases como as seguintes, assumindo que os nomes de pessoas, bem como os cargos, são correctamente classificados, o resultado final será diferente para cada uma delas.

Exemplo 9:

...foram convidados: Cavaco Silva, Presidente da República, José Sócrates, Primeiro Ministro, Alberto Costa, Ministro da Justiça,...

...estiveram presentes: o Presidente da República, Cavaco Silva, o Primeiro Ministro, José Sócrates, o Ministro da Justiça, Alberto Costa,... ■

A aplicação da regra na primeira frase gera as relações certas: temCargo(Cavaco Silva, Presidente da República), temCargo(José Sócrates, Primeiro Ministro) e temCargo(Alberto Costa, Ministro da Justiça). No entanto, na segunda frase, as relações criadas são outras: temCargo(Cavaco Silva, Primeiro Ministro), temCargo(José Sócrates, Ministro da Justiça).

A recolha destas relações gera incoerências e provoca o armazenamento de informação incorrecta na base de dados. Para contornar este problema, o sistema assume, devido à dimensão do *corpus*, que as relações se encontram disponíveis sob outras formas (que não em lista) com bastante maior frequência e que, dessa forma, as relações certas serão armazenadas em base de dados em maior quantidade do que as relações erradas.

3.5 Sumário

Baseando o seu processamento e a sua acção em quantidades de informação que atingem a dimensão de vários *MegaBytes*, a arrumação dos dados para posterior acesso pelo sistema revela-se de grande importância.

Este capítulo fez referência ao modo como a informação pertencente ao *corpus* foi guardada, de forma a ser posteriormente utilizada. Apresentaram-se as estruturas das várias bases de dados, revelando-se o objectivo da sua criação e a forma como foram criadas, recorrendo às ferramentas de PLN, tendo sempre em mente o objectivo inerente ao sistema: responder a perguntas.

4 Procura da Resposta

Qual a cor do cavalo branco de Napoleão?

– Adivinha Portuguesa

4.1 Introdução

A procura da resposta é uma tarefa num sistema de QA que combina extracção de informação, processamento de língua natural, análise e restringimento da informação a devolver ao utilizador. Um dos obstáculos inerentes a esta tarefa relaciona-se com o facto da formulação linguística da pergunta submetida pelo utilizador ser, em bastantes casos, muito diferente da maneira como a informação aparece no *corpus*. Considere-se, a título de exemplo, a pergunta: "Em que ano se afundou o Titanic?" Recorrendo ao CETEMPúblico (Rocha & Santos, 2000; Santos & Rocha, 2001) [1] verifica-se que 651 frases contêm a palavra Titanic. Dentro destas, apenas 15 contêm a palavra afundou e 14 a palavra afundar.

Atente-se, também, nos dois excertos seguintes retirados do mesmo *corpus*:

Na viagem inaugural para Nova Iorque, a 14 de Abril de 1912, o Titanic desceu para o leito de morte a três mil metros de profundidade...

Não localizou apenas o «Titanic» a 1 de Setembro de 1985, sepultado desde a noite de 15 de Abril de 1912 ao largo da Terra Nova..

Qualquer uma das frases contém a resposta certa à pergunta; implicam, porém, um nível de processamento linguístico aprofundado, não apenas assente na semelhança entre termos presentes na pergunta e no *corpus*.

Neste capítulo é descrito o processamento efectuado pelo sistema na procura da resposta a uma pergunta. Na secção 4.2 apresenta-se o módulo de análise e interpretação da pergunta submetida; na secção 4.3 referem-se e discutem-se cada uma das estratégias utilizadas pelo sistema para extracção da resposta final; finalmente, na secção 4.4 descreve-se o mecanismo de relaxamento de restrições implementado pelo sistema e que lhe permite fazer uso de um conjunto dessas estratégias para responder a uma pergunta.

[1] disponível *online* em `http://acdc.linguateca.pt/cetempublico/acesso_CP.html`

4.2 Análise e Interpretação da Pergunta

A fase de análise e interpretação da pergunta tem como objectivo a extracção de informação da pergunta efectuada. É responsável, também, pelo encaminhamento dessa informação para o script que tratará de responder a essa pergunta submetida. Sendo um módulo constituinte do QA@L^{2}F, esta secção dedica-se à sua descrição; sai, no entanto, do âmbito do trabalho desenvolvido, pelo que não se irão apresentar em detalhe as funcionalidades desta fase de análise e interpretação da pergunta.

O construtor de *frames* é responsável por identificar:

- o *script* a ser chamado, dependendo da pergunta efectuada;
- a entidade-alvo da pergunta (como uma pessoa ou uma cidade);
- todas as entidades mencionadas e palavras auxiliares encontradas na pergunta.

Considere-se, a título de exemplo, a pergunta "Quem é Boaventura Kloppenburg?". O ficheiro XML resultante da cadeia de PLN (descrita na secção 1.3.2) contém (entre outras) a seguinte informação:

Ficheiro de exemplo 4:

```
<NODE num="11" tag="NOUN" start="8" end="29">
<FEATURE attribute="PEOPLE" value="+"/> (1)
<NODE num="4" tag="VERB" start="8" end="17">
Boaventura
</NODE>
<NODE num="6" tag="NOUN" start="19" end="29">
Kloppenburg
</NODE>
</NODE>
...
<DEPENDENCY name="TARGET_WHO_PEOPLE"> (2)
<PARAMETER ind="0" num=``11" word="Boaventura Kloppenburg"/>
</DEPENDENCY>
```

■

A dependência TARGET_WHO_PEOPLE é identificada (em **(2)**), bem como a entidade mencionada do tipo PEOPLE (em **(1)**). De referir que cada dependência tem correspondência directa com um *script* de extracção de resposta. Neste sentido, é a existência de uma determinada dependência no

ficheiro XML resultante do processamento linguístico da pergunta que vai determinar qual o *script* a ser chamado.

A *frame* seguinte é construída a partir do ficheiro anterior:

Frame 1: Quem é Boaventura Kloppenburg?

```
SCRIPT script-who-people.pl
TARGET "Boaventura Kloppenburg"
ENTIDADES "Boaventura Kloppenburg " PEOPLE
```

■

A *frame* indica que: o *script* a ser chamado é o `script-who-people.pl`, a entidade-alvo é "Boaventura Kloppenburg" e foi descoberta na pergunta uma entidade mencionada do tipo PEOPLE: "Boaventura Kloppenburg". O nome do *script* indica explicitamente qual o tipo de pergunta formulada.

Noutro exemplo, a pergunta "Onde fica o parque Eduardo VII?" corresponde ao excerto de ficheiro XML seguinte:

Ficheiro de exemplo 5:

```
<NODE num="8" tag="NP" start="20" end="26">
<FEATURE attribute="PEOPLE" value="+"/> (1)
<TOKEN pos="NOUN" start="20" end="26">
Eduardo
<READING lemma="Eduardo" pos="NOUN">
</READING>
</TOKEN>
</NODE>
...
<NODE num="10" tag="NUM" start="28" end="30">
<FEATURE attribute="NUM" value="+"/> (2)
<TOKEN pos="NUM" start="28" end="30">
VII
<READING lemma="vii" pos="NUM">
</READING>
</TOKEN>
</NODE>
...
<DEPENDENCY name="TARGET_WHERE_1"> (3)
```

```
<PARAMETER ind="0" num="6" word="parque"/>
</DEPENDENCY>
<DEPENDENCY name="EXTRA"> (4)
<PARAMETER ind="0" num="18" word="Eduardo"/>
</DEPENDENCY>
<DEPENDENCY name="EXTRA"> (5)
<PARAMETER ind="0" num="20" word="VII"/>
</DEPENDENCY>
```
■

Neste pedaço de ficheiro é visível a existência de duas entidades mencionadas (em **(1)** e **(2)**), bem como uma dependência do tipo TARGET_WHERE_1 (em **(3)**). Um novo tipo de dependência teve de ser criada, a dependência com o nome EXTRA (em **(4)** e **(5)**). Com a utilização desta dependência, a entidade-alvo da pergunta passa a ser o resultado da concatenação da entidade-alvo definida na dependência TARGET_WHERE_1, com conceitos capturados nas dependências TARGET, ou seja, parque Eduardo VII, ao invés de apenas parque. Este tipo de dependências surgiu para colmatar o facto de alguns conceitos não serem capturados como um todo pela cadeia de PLN (numa entidade mencionada, ou num único nó da árvore sintáctica), tal como se verifica no exemplo anterior.

A *frame* construída, responsável por chamar o `script-where-basic.pl`, com a entidade-alvo "parque Eduardo VII " e as entidades mencionadas descobertas na pergunta, "Eduardo ", com o tipo PEOPLE, e "VII ", com o tipo NUM, é a seguinte:

Frame 2: Onde fica o parque Eduardo VII?

```
SCRIPT script-where-basic.pl
TARGET "parque Eduardo VII "
ENTIDADES "Eduardo " PEOPLE "VII " NUM
AUXILIARES
```
■

De notar que esta aproximação, apesar de recolher conceitos chave da pergunta, como as suas entidades mencionadas, e o tipo de resposta que se pretende, é considerada limitada. Por um lado, não tem em conta expressões temporais. Conceitos como "antes", "depois" ou "entre <data1> e <data2>" não são tratados e, como tal, não possuem qualquer significado para o sistema. Tratam-se apenas de palavras auxiliares, sobre as quais o sistema não efectua qualquer processamento especial. Por outro, a abordagem poderia também ser estendida de maneira a fornecer pistas referentes à categoria sintáctica onde a resposta se pode encontrar no *corpus* (no complemento directo, circunstancial de lugar ou de tempo de um verbo, por exemplo).

4.3 Extracção da Resposta Final

A etapa que procede à extracção da resposta final é a última na cadeia de processamento do sistema de QA. Compreende os passos efectuados pelo sistema, já na posse de toda a informação recolhida da pergunta na fase anterior de análise e interpretação, por forma a devolver uma resposta final à questão efectuada.

Dependendo quer da pergunta recebida como entrada, quer da maior ou menor facilidade em encontrar a resposta, o sistema tem ao seu dispor um conjunto de estratégias e alternativas.

As secções seguintes apresentam em detalhe cada uma dessas estratégias desenvolvidas, que permitem ao sistema atingir o seu objectivo de responder de forma correcta e precisa a cada pergunta. Para cada uma das estratégias, são ainda evidenciados os seus pontos fortes e fracos, bem como o tipo de perguntas a que melhor estão habilitadas a responder.

4.3.1 Emparelhamento de Padrões Linguísticos

Nas perguntas directas o sistema interroga a tabela de factos, na base de dados, do tipo correspondente. Uma pergunta do tipo "Quem é Mário Soares?" possibilita a construção da seguinte *frame*:

Frame 3: Quem é Mário Soares?

```
SCRIPT script-who-people.pl
TARGET "Mário Soares"
ENTIDADES "Mário Soares " PEOPLE
```
■

Esta informação, proveniente da fase de análise e interpretação da pergunta, é traduzida, posteriormente, para a seguinte instrução MySQL:

Interrogação 4: Quem é Mário Soares?

```
SELECT title, id, confidence, count, CHAR_LENGTH(title) as maxChar
FROM FACT_PEOPLE
WHERE name="Mário Soares"
GROUP BY confidence DESC, count DESC, maxChar DESC;
```
■

O sistema recolhe, por ordem decrescente de confiança no padrão, de número de vezes em que aparecem no corpus e de número de caracteres, os campos `title` que correspondem a entradas na

tabela com o campo `name` igual a Mário Soares.

O número de caracteres tem relevância nesta interrogação na medida em que uma resposta contendo informação a mais é preferida, pelo sistema, a uma resposta que contém informação insuficiente. Considere-se, por exemplo, a informação presente nas bases de dados com respeito a Kim Il Sung, e visível na tabela 4.1.

FACT_PEOPLE

id	name	title	confidence	count
970	Kim Il Sung	líder da Coreia do Norte	50	1
434	Kim Il Sung	presidente	50	2
87	Kim Il Sung	Presidente da Coreia do Norte	50	2
24	Kim Il Sung	Presidente norte-coreano	50	1

Tabela 4.1: Entradas na tabela *FACT_PEOPLE* contendo informação acerca de Kim Il Sung.

Como ambos têm o mesmo grau de confiança associado e a sua contagem no *corpus* é igual, ambos presidente e Presidente da Coreia do Norte seriam passíveis de ser escolhidos como resposta a uma eventual pergunta acerca do cargo de Kim Il Sung.

Este tipo de abordagem evita também que a uma pessoa seja associada a abreviatura do seu cargo (como num caso semelhante ao da tabela 4.2), tal acontecendo apenas no caso limite em que mais nenhuma relação entre a pessoa e o seu cargo seja recolhido do *corpus*.

FACT_PEOPLE

id	name	title	confidence	count
621	Anibal Cavaco Silva	Prof .	99	2
2334	Anibal Cavaco Silva	professor	99	2

Tabela 4.2: Entradas na tabela *FACT_PEOPLE* contendo informação acerca de Cavaco Silva.

Se a interrogação à base de dados devolver uma ou mais entradas da tabela, o sistema assume que a entrada devolvida em primeiro lugar é a resposta à pergunta, escolhendo sempre essa.

Vantagens da Estratégia de Emparelhamento de Padrões Linguísticos

Esta estratégia apresenta como vantagens:

- permitir responder a um alargado conjunto de perguntas directas;
- não exigir qualquer processamento linguístico após a submissão da pergunta ao sistema, sendo toda a computação feita *a priori*.

Desvantagens da Estratégia de Emparelhamento de Padrões Linguísticos

As desvantagens inerentes a esta abordagem são:

- a dependência da informação presente nos textos jornalísticos. Os textos jornalísticos são, eles próprios, bastante influenciados pela forma como as pessoas falam na sua vida quotidiana (porque é para elas que os jornais estão direccionados, e não para fazerem parte do *corpus* de um sistema de QA). Considere-se o exemplo de Cavaco Silva (tabela 4.3). Apesar de ministro [2] de Portugal e sendo este, possivelmente, o seu cargo de maior relevo para dar resposta a uma pergunta do tipo "Quem é Cavaco Silva?", é muito mais frequente no *corpus* a relação entre Cavaco Silva e o cargo Professor.

FACT_PEOPLE

id	name	title	confidence	count
569	Cavaco Silva	ministro	99	3
1339	Cavaco Silva	ministro	99	4
161	Cavaco Silva	Prof .	99	9
52	Cavaco Silva	professor	99	22

Tabela 4.3: Entradas na tabela *FACT_PEOPLE* contendo informação acerca de Cavaco Silva.

- o facto de não se saber até que ponto a informação contida na base de dados corresponde, de facto, a uma resposta, apesar do sistema confiar na informação que tem armazenada. Veja-se o exemplo da tabela 4.4. Será que, quer major quer presidente, são, qualquer uma delas, uma boa e útil resposta a dar a uma pergunta como "Quem é Valentim Loureiro?".

FACT_PEOPLE

id	name	title	confidence	count
11	Valentim Loureiro	major	99	8
243	Valentim Loureiro	presidente	99	2

Tabela 4.4: Entradas na tabela *FACT_PEOPLE* contendo informação acerca de Valentim Loureiro

4.3.2 Reordenação de Formulações Linguísticas

A estratégia baseada na reordenação de formulações linguísticas aplica-se para responder quer a perguntas de definição (como "Quem foi Ésquilo?" ou "O que é um barrete frígio?"), quer a perguntas cuja

[2] De referir ainda, que, à altura destes textos, Cavaco Silva era primeiro ministro de Portugal. As regras de dependência não conseguiram, porém, capturar este cargo, já que apenas a palavra ministro é identificada como entidade mencionada TITLE, sendo que as palavras primeiro ministro não são consideradas como tal.

resposta esperada seja uma lista (como "Diga uma escritora sarda." ou "Quero o nome de um vinho húngaro."). [3] O sistema utiliza a Wikipedia por forma a responder a estes dois tipos de questões.

Em primeiro lugar, a fase de análise e interpretação da pergunta é responsável por identificar e extrair o conceito principal presente na questão. Atendendo às perguntas dadas como exemplo no parágrafo anterior, estes seriam: Ésquilo, barrete frígio, escritora sarda e vinho húngaro.

O sistema prossegue, então, com uma procura nos artigos da Wikipedia armazenados em base de dados. A abordagem seguida difere, no entanto, para cada um dos tipos de questões.

Nas perguntas de definição, o sistema tira proveito do formato da Wikipedia, em que a informação textual presente num artigo diz directamente respeito ao título do mesmo artigo. Desta forma, o sistema interroga a base de dados pelo artigo da Wikipedia cujo título é o conceito procurado. Assim, se o conceito for Ésquilo, é utilizada a seguinte interrogação:

Interrogação 5: Quem foi Ésquilo?

```
SELECT page_text, page_id
FROM PAGE_TEXT
WHERE page_title REGEXP "Ésquilo";
```
■

Na posse do artigo da Wikipedia relacionado com o conceito procurado, o sistema analisa a informação textual contida apenas na sua primeira frase. Após análise dos textos da Wikipedia, verificou-se que a resposta a este tipo de questões se encontra frequentemente apenas na primeira frase do artigo, não existindo necessidade de alargar o processamento a todo o artigo. Nessa primeira frase, o sistema efectua uma pesquisa por padrões que permitam responder à pergunta formulada. Estes padrões correspondem a formulações linguísticas compostas pelo conceito procurado seguido (imediatamente ou não) pelo verbo "ser" conjugado numa das seguintes formas:

- na terceira pessoa do singular, no tempo presente do indicativo, por exemplo, "...Ésquilo foi...";
- na terceira pessoa do singular, no tempo passado perfeito do indicativo, por exemplo, "...barrete frígio é...";

Considere-se a informação contida na base de dados com informação acerca de Ésquilo e apresentada na tabela 4.5. O padrão procurado é encontrado no artigo e a resposta dada pelo sistema é: "um poeta trágico grego".

[3] Apesar de não serem consideradas, no âmbito do CLEF, como perguntas do tipo *List*, o sistema assume que este tipo de perguntas espera como resposta uma lista de um só elemento, utilizando esta estratégia.

WIKIPEDIA

id	page_title	page_text
45162	Ésquilo	Ésquilo (Elêusis c. 525 a.C. - Gela 456 a.C.) foi um poeta trágico grego. É considerado como o fundador da tragédia. Foi soldado em Maratona, Salamina e Plateias (o que explica várias peças de cariz militarista,...

Tabela 4.5: Entrada na base de dados da Wikipedia contendo informação acerca de Ésquilo.

De referir que esta abordagem restringe a resposta devolvida ao utilizador à informação essencial, evitando qualquer tipo de informação supérflua: por um lado, permite a existência de informação não essencial para a resposta (como "(Elêusis c. 525 a.C. - Gela 456 a.C.)" presente no exemplo anterior) no padrão pesquisado, não contando com esta informação para a resposta final; por outro, depois de encontrar o padrão, apenas devolve a definição do conceito até encontrar ou um ponto final ou uma vírgula.

Nos casos das perguntas que esperam como resposta uma lista, o sistema procede de forma inversa. Começa por interrogar a base de dados por artigos que no seu texto contenham a conceito procurado. No caso da pergunta "Diga uma escritora sarda.", a seguinte interrogação *full-text* é criada:

Interrogação 6: Mencione uma escritora sarda.

```
SELECT DISTINCT page_text, page_title
FROM PAGE_TEXT WHERE MATCH(page_text) AGAINST ("escritora sarda") LIMIT 25
```
■

Na posse dos 25 artigos recolhidos pela interrogação, o sistema faz uma procura pelo conceito "escritora sarda". Nesta abordagem, o sistema assume, se esta procura tiver resultado, que existe uma elevada probabilidade de que o conceito faça parte de uma definição dada pela Wikipedia sendo o título do artigo a resposta à pergunta. No exemplo fornecido, o conceito "escritora sarda" é composto por 2 palavras. Caso a procura não tenha dado resultado, e com o intuito de considerar, também, as frases em que o conceito não se encontra escrito exactamente como esperado, é feita uma pesquisa pelas palavras que compõem o conceito, em separado ("escritora" e "sarda"). A tabela 4.6 mostra a entrada na tabela referente a Grazia Deledda em que tal acontece e da qual se pode extrair a resposta à pergunta "Diga uma escritora sarda." utilizando esta estratégia.

No caso destas perguntas que esperam um determinado número de respostas, por exemplo "Diga três livros de José Saramago.", a procura referida nos dois últimos parágrafos é efectuada esse número de vezes (três).

WIKIPEDIA

id	page_title	page_text
265412	Grazia_Deledda	Grazia Deledda (Nuoro, 27 de setembro de 1871 — Roma, 15 de agosto de 1936) foi uma escritora e poeta sarda, vencedora do Prémio Nobel de Literatura de 1926...

Tabela 4.6: Entrada na base de dados da Wikipedia contendo informação acerca de Grazia Deledda.

Vantagens da Estratégia de Reordenação de Formulações Linguísticas

Esta estratégia tem como vantagem:

- utilizar a Wikipedia como fonte de informação. Devido ao formato particular em que os seus textos se encontram escritos, facilita a extracção das respostas às perguntas dos tipos definição e lista.

Desvantagens da Estratégia de Reordenação de Formulações Linguísticas

A desvantagem atribuída a esta estratégia é:

- a natureza rudimentar dos padrões utilizados, que não permite restringir a informação devolvida ao utilizador àquela considerada essencial.

  Seja a pergunta "Quem é Aníbal Cavaco Silva?". Utilizando a aproximação baseada na reordenação de formulações linguísticas, a resposta devolvida pelo sistema é: "Presidente da República PortuguesaAníbal António Cavaco Silva, GCC (Boliqueime, Loulé, 15 de Julho de 1939) é o actual Presidente da República Portuguesa.". Esta situação verifica-se já que o conceito procurado, Aníbal Cavaco Silva, não se encontra, de facto, nesta frase. O que existe nesta frase é o conceito Aníbal António Cavaco Silva, diferente do anterior. Como tal, o sistema não consegue fazer o emparelhamento com o padrão esperado e retorna a primeira frase completa do artigo. Apesar de conter informação a mais do que é realmente necessário ("o actual Presidente da República Portuguesa" seria uma resposta mais precisa), o sistema devolve a resposta correcta.

4.3.3 Emparelhamento de Entidades Mencionadas

O sistema de QA, na posse de toda a informação recolhida da pergunta, faz uma pesquisa nas tabelas contendo entidades mencionadas.

A ideia inerente a esta estratégia é fazer um emparelhamento entre as entidades mencionadas encontradas na pergunta e as entidades mencionadas encontradas nos parágrafos do *corpus* (e previamente

guardados na base de dados na tabela *SNIPPET*). O sistema assume que a resposta estará definida num parágrafo que contenha as mesmas entidades mencionadas encontradas na pergunta. Depois de reduzir o *corpus* a este conjunto de parágrafos, o sistema procura a entidade mencionada mais frequente do tipo requerido e devolve-a como resposta final à pergunta.

Considere-se, por exemplo, a pergunta "Quem abriu os Jogos Olímpicos de 1948?". Da fase de análise e interpretação da pergunta submetida, resulta a seguinte *frame*:

Frame 4: Quem abriu os Jogos Olímpicos de 1948?

```
SCRIPT script-who-v-geral.pl
TARGET VAZIO
ENTIDADES "Jogos Olímpicos " EVENT "1948 " NUM
AUXILIARES "abriu" ACTION "os Jogos Olímpicos"
```
■

O *script* chamado, que pretende encontrar como resposta o nome de uma pessoa (entidade mencionada do tipo PEOPLE), lidará com uma entidade-alvo vazia, com as entidades mencionadas "Jogos Olímpicos " (EVENT) e "1948 " (NUM).

As entidades, do tipo NUM, entre 1500 e 2500 são consideradas como entidades mencionadas do tipo TIME. Assim sendo, o sistema pesquisa nas tabelas *NE_TIME* e *NE_EVENT* pelo identificador das duas entidades mencionadas e faz a sua fusão nos parágrafos do *corpus* recorrendo à seguinte interrogação à base de dados[4]:

Interrogação 7: Quem abriu os Jogos Olímpicos de 1948?

```
SELECT event0.snippetId
FROM SNIPPET_HAS_NE_EVENT as event0, SNIPPET_HAS_NE_TIME as time2
WHERE event0.neEventId=51
and time2.neTimeId=334
and event0.snippetId=time2.snippetId;
```
■

O esquema da figura 4.1 representa a fase de fusão efectuada pelo sistema em que as entradas na tabela *SNIPPET* da base de dados contendo as entidades mencionadas "1948" e "Jogos Olímpicos" são seleccionadas.

Após a recolha dos parágrafos onde se pensa estar resposta, o número de entidades mencionadas

[4]O identificador descoberto, na tabela *NE_TIME*, da entidade mencionada do tipo TIME ("1948") é 334; o identificador descoberto, na tabela *NE_EVENT*, da entidade mencionada do tipo EVENT ("Jogos Olímpicos") é 51.

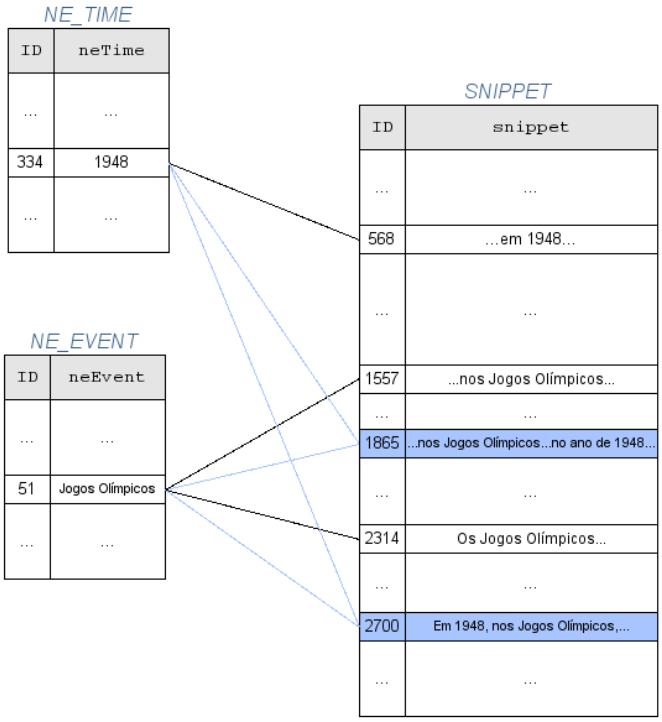

Figura 4.1: Fusão entre as entidades mencionadas presentes no *corpus* e na pergunta.

do tipo esperado pela pergunta submetida é contado, sendo devolvida a entidade mencionada mais frequentemente. A figura 4.2 representa esquematicamente a fase de escolha da resposta final desta estratégia. Neste caso, a resposta escolhida seria "Rei Jorge VI".

Vantagens da Estratégia de Emparelhamento de Entidades Mencionadas

A estratégia descrita tem como vantagem:

- assumir que a resposta esperada para uma pergunta será uma entidade mencionada. Nesse sentido, se se tiver em consideração que grande parte do espectro de perguntas colocadas ao sistema pertencerá aos tipos `Quem...?`,`Onde...?`, `Quando...?` e `Quanto...?`. Basta analisar, por exemplo, o conjunto de perguntas realizadas no CLEF no ano de 2007 (disponíveis no apêndice A) para se verificar que, em grande parte delas, se espera como resposta uma entidade mencionada dos tipos PEOPLE, LOCATION, TIME, entre outras.

Desvantagens da Estratégia de Emparelhamento de Entidades Mencionadas

As desvantagens desta estratégia são:

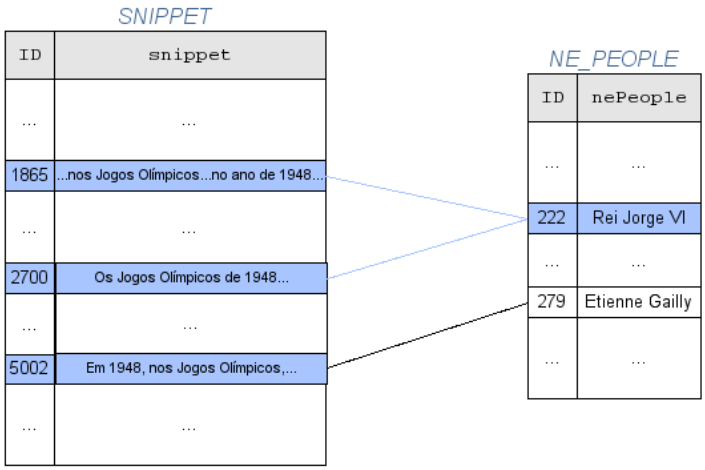

Figura 4.2: Escolha da resposta final.

- as perguntas do tipo `Como...?` ou `Porque...?`, não serem solucionadas utilizando esta abordagem, já que não esperam uma entidade mencionada como resposta, necessitando de outro tipo de processamento;
- a recolha de entidades mencionadas na questão submetida pode não ser uma tarefa trivial. Considerem-se as perguntas: "Quantos focos tem uma elipse?" e "Quem era rei de Portugal em 1860?". Apesar de na segunda poderem ser reconhecidas como entidades mencionadas os conceitos: rei (TITLE), Portugal (LOCATION) e 1860 (TIME), na primeira nenhuma entidade mencionada será reconhecida pelo sistema;
- ser necessária a existência de um emparelhamento entre as entidades mencionadas na pergunta e no corpus o que, por vezes, não acontece;
- existir uma dependência em relação ao tamanho de cada pedaço de informação presente em base de dados. Um dos impactos negativos, relacionado directamente com a forma como a informação foi organizada e armazenada em base de dados (ver secção 3.3 deste documento), reside sobretudo no facto desta estratégia ter impedido as ferramentas de PLN de serem utilizadas com o máximo de eficiência na altura do pré-processamento do *corpus*.

4.3.4 *Brute Force* com Pós-Processamento de Língua Natural

Se nenhuma das estratégias referidas nas secções anteriores tiver retornado uma possível resposta correcta à pergunta formulada, o sistema QA@L^{2}F tenta encontrá-la recorrendo a uma estratégia *brute force* com pós-processamento de língua natural. Assim, é efectuada uma interrogação *full-text* à base de

dados contendo o *corpus* não processado, utilizando toda a informação recolhida na fase de análise e interpretação da pergunta. O processamento de língua natural é efectuado *a posteriori*, nos parágrafos melhor classificados e devolvidos pela interrogação nos primeiros lugares. Este processamento visa a descoberta de entidades mencionadas de categoria idêntica ao tipo de resposta requerida pela pergunta efectuada. Assim, uma pergunta `Quem...?` espera como resposta uma entidade mencionada categorizada como people ou um title; uma pergunta `Onde...?` espera uma location;...

Após a identificação das entidades mencionadas nos parágrafos recolhidos da base de dados, aquela que tiver o maior número de ocorrências é considerada, pelo sistema, como sendo a resposta final à pergunta.

Vantagens da Estratégia *Brute Force* com Pós-Processamento de Língua Natural

A vantagem desta estratégia é:

- considerar todo o *corpus* jornalístico, e não apenas aquele que foi pré-processado. Num caso em que o *corpus* é de elevada dimensão (como acontece num sistema de QA), esta estratégia pode significar a independência relativamente às restrições de tempo de pré-processamento que o sistema tem. Quando não existe tempo para pré-processar, ou quando este se revela insuficiente, esta estratégia efectua uma pesquisa em todo o corpus, sendo o processamento linguístico efectuado depois da submissão da pergunta.

Desvantagens da Estratégia *Brute Force* com Pós-Processamento de Língua Natural

As desvantagens relativas a esta estratégia são:

- não se conseguir determinar à partida o número de parágrafos que devem ser retornados da base de dados pela interrogação *full-text*. Por um lado, se for demasiado baixo, corre-se o risco de nenhum deles conter o termo que esperamos encontrar; por outro, se for demasiado alto, os termos encontrados que emparelham com o resultado pretendido poderão levar a que se escolha o termo incorrecto, se se encontrarem em maior número.

  Por exemplo, considere-se que o número de parágrafos devolvidos é 5 e que a pergunta a que queremos responder é "Quem é Boris Becker?".

Corpus 3:

A volta de Boris Becker aos top 3 tem seu lado negativo. Dá munição à prepotência do técnico Nick Bollettieri. O fanfarrão declarou à revista alemã Stern que é o maior responsável pela recuperação de Becker. Fiz profundas mudanças técnicas para adaptar o tênis de Boris (..)

Outro que vai a Austrália é o alemão Boris Becker. Sua intenção, aliás, é colocar em apuros o número 1 do mundo. Becker ganhou o torneio em 91.

Para o [Frankfurter Allgemeine Zeitung], pouco dado a sensacionalismos, a vitória de Schumacher é comparável à vitória de Boris Becker no torneio de Wimbledon, em 1985, quando a Alemanha descobriu o seu primeiro grande campeão no ténis. O [Tageszeitung] de Berlim utiliza a mesma comparação: [Steffi [Graff] e Boris [Becker] (...)

A vitória arrasadora de Boris Becker no ATP Tour de Los Angeles no último fim-de-semana foi recebida com vibração pela imprensa especializada. Segundo relato de comentarista argentino, Becker(...)

Com tudo isso, Agassi mantém sua figura na imprensa e dá retorno aos patrocinadores. Boris Becker é outro que forjou uma imagem. Suas declarações polêmicas e seu não menos polêmico casamento com a negra Barbara Feltus garantiram manchetes a ele mesmo quando despencou no ranking. Stich venceu sete torneios em 93, contra apenas um de Becker. Com publicidade, Becker faturou (...)

Todos os parágrafos mencionam e fazem referência ao ténis, desporto praticado por Boris Becker, mas nenhum deles contém o cargo tenista. Assim, o único cargo existente em todos eles é técnico (na segunda frase do primeiro parágrafo), sendo que o sistema o irá devolver, porém erradamente, como resposta à pergunta submetida.

- haver baixa reactividade na interacção com o utilizador. Num sistema de QA que se queira reactivo, esta abordagem poderá não ser viável: o processamento de todo o *corpus* é efectuado depois da submissão da pergunta, aumentando assim o tempo de resposta ao utilizador (encontrando-se na ordem dos 3 minutos). A questão que se coloca é, neste caso, quanto tempo está um utilizador disposto a esperar pela resposta à sua pergunta?

4.4 Mecanismo de Relaxamento de Restrições

O sistema QA@L^{2}F faz uso de um mecanismo de relaxamento de restrições, utilizando as estratégias descritas na secção 4.2 para extracção da resposta final a uma pergunta. A ideia subjacente a este mecanismo é aplicar um conjunto de estratégias sucessivas, começando com a que melhor se aplica a um tipo de pergunta e seguindo um caminho de estratégias, cada vez mais flexíveis em termos de restrições, até à descoberta da resposta final. Dependendo do tipo de pergunta recebida, o caminho tomado é diferente, tendo sido, no entanto, determinado à partida.

Considere-se, por exemplo a pergunta "Quem escreveu ≪Os Lusíadas≫?". A primeira estratégia que o sistema aplica é a de emparelhamento de padrões linguísticos: o sistema procura na base de dados factual o autor para o qual foi encontrado um padrão da categoria CULTURE que o relacione com ≪Os Lusíadas≫.

Caso esta estratégia falhe devido à ausência na base de dados de uma relação entre a obra e o seu autor, o sistema recorre à estratégia de emparelhamento de entidades mencionadas. Com esta estratégia o sistema tenta encontrar o nome de pessoa mais frequente nos parágrafos em que a entidade mencionada do tipo CULTURE ≪Os Lusíadas≫ também se encontra.

Novamente, em caso de insucesso, o sistema utiliza a estratégia *brute force* com pós-processamento de língua natural, efectuando uma interrogação *full-text* à base de dados contendo o *corpus* não processado pelo conceito ≪Os Lusíadas≫. Nos dez parágrafos melhor classificados faz uma procura quer pelo nome de pessoa, quer pelo nome próprio, mais frequentes.

Se a estratégia *brute force* não encontrar resultados, o sistema assume que a resposta não se encontra no *corpus* e devolve NIL como resposta.

Script	Resposta	Estratégia	Próximo *Script*
`who-people`	Cargo	Padrões Linguísticos	`who-people-wiki`
`who-people-wiki`	Cargo	Reord. Form. Linguísticas	`who-chaos-people`
`where-basic`	Localizacao	Padrões Linguísticos	`where-chaos-location`
`where-chaos-location`	Localizacao	*Brute Force*	
`where-v`	Localizacao	Emp. Ent. Mencionadas	`where-chaos`
`where-chaos`	Localizacao	*Brute Force*	
`what-abbr`	Abreviatura	Padrões Linguísticos	`what-abbr-wiki`
`what-abbr-wiki`	Abreviatura	Reord. Form. Linguísticas	`what-def-chaos`
`what-def`	Definição	Padrões Linguísticos	`what-def-chaos`
`what-def-chaos`	Definição	*Brute Force* + Reord. Form. Linguísticas	
`list`	Lista	Reord. Form. Linguísticas	

Tabela 4.7: *Scripts* para extracção da resposta final.

A tabela 4.7 apresenta os nomes de alguns *scripts*, desenvolvidos na linguagem Perl, utilizados pelo sistema para extracção da resposta final. Para cada um deles, são indicados: o tipo de resposta esperada, a estratégia implementada e o próximo *script* a ser chamado caso não tenha sido encontrada resposta à

pergunta submetida. Desta forma, ilustram-se os caminhos possíveis e que compõem o mecanismo de relaxamento de restrições do sistema.

4.5 Sumário

O capítulo que aqui termina abordou a tarefa de procura da resposta. Esta tarefa começa com a fase de análise e interpretação da pergunta, responsável por recolher toda a informação importante existente na pergunta e encaminhá-la para a fase seguinte de extracção da resposta final. Esta fase, por sua vez, tem como função devolver a resposta correcta à pergunta efectuada. Para este efeito, tem à sua disposição quatro estratégias que utiliza em função da pergunta submetida. Estas podem ser aplicadas sucessivamente para uma mesma pergunta, caso a estratégia que melhor se aplique tenha falhado na tentativa de encontrar a resposta, constituindo este o mecanismo de relaxamento de restrições implementado pelo sistema.

5 Avaliação

It is not every question that deserves an answer.

– Publilius Syrus

5.1 Introdução

Constituindo o objectivo de base do projecto Clefomania, depois de implementada a arquitectura geral do sistema, o desenvolvimento do QA@L^{2}F culminou com a sua avaliação no CLEF por forma a testar a sua eficiência.

Foi também realizada uma segunda avaliação, desta vez no laboratório. Na posse dos resultados obtidos pelo sistema na avaliação anterior e, depois da sua análise e interpretação, foram feitas alterações superficiais ao sistema (não implementando, porém, nenhuma outra estratégia ou fazendo alterações profundas).

Este capítulo analisou a avaliação realizada ao sistema: a secção 5.2 apresenta os resultados no contexto do CLEF e a secção 5.3 descreve os resultados obtidos no segundo momento de avaliação, feita no L^{2}F.

5.2 Avaliação no CLEF 2007

O sistema QA@L^{2}F foi avaliado no âmbito do fórum CLEF (Mendes et al., 2007) através da submissão ao sistema de um conjunto de 200 perguntas que, tal como em anos anteriores (ver secção 2.2.3 deste documento), podem:

- ter uma das categorias *Factoid* (F), *Definition* (D) ou *List* (L). Seguem-se exemplos de uma pergunta para cada uma dessas categorias:

  Pergunta F: Quantas regiões tem Marrocos?

  Pergunta D: O que é a constante de Néper?

  Pergunta L: Quais são as cidades-estado da Alemanha? ■

- ter como resposta a indicação de que a pergunta não tem resposta (*NIL*);

- conter restrições temporais (*Temporaly Restricted*). Por exemplo:

  Pergunta: Qual era a divisa austríaca antes de 2002? ■

Na edição de 2007, algumas perguntas foram organizadas em grupos, por forma a testar a capacidade de resolução de anáforas e elipses dos sistemas. Por exemplo:

Grupo 1940: O que é o Tux?

Grupo 1940: Que animal é?

Grupo 1940: Quem o criou?

Grupo 1940: Quando? ■

A tabela 5.1 mostra o número de perguntas existentes para cada uma das características descritas e a sua relação com o número total de perguntas. De notar que, no conjunto de perguntas pertencentes a um grupo (75 perguntas), se encontram destacadas aquelas que necessitam de tratamento anáforico/elíptico para serem respondidas (50 perguntas). [1]

	# de Perguntas	**% de Perguntas**
Factoid	159	79.50
Definition	31	15.50
List	10	5.00
Temporaly Restricted	*19*	*9.50*
Pertence a um grupo	75	37.50
Anáfora/Elipse	*50*	*25.00*
Não pertence a um grupo	125	62.50
Total	200	100.00

Tabela 5.1: Resumo das perguntas submetidas ao sistema na sua avaliação.

Cada sistema possui um total de duas submissões para avaliar o seu sistema. Isto é, no máximo poderão ser efectuadas duas avaliações ao sistema para o mesmo conjunto de 200 perguntas. A resposta dada pelo sistema a cada pergunta deve ser única (no caso das perguntas da categoria *List* deve ser uma lista contendo todos os itens necessários) e deve ser acompanhada pelo parágrafo do texto que a sustenta, bem como o identificador do documento a que pertence.

O conjunto de perguntas submetidas ao sistema estão disponíveis neste documento e podem ser consultadas no apêndice A.

5.2.1 Medidas e Métricas de Desempenho

Cada resposta é avaliada e pontuada de acordo com as seguintes medidas:

[1] Apenas a primeira pergunta do grupo não necessita de tratamento anafórico/elíptico; as restantes, porém, por estarem directamente relacionadas com as anteriores do grupo, precisam deste tipo de processamento.

- **R** *(Right)*: se está correcta;
- **W** *(Wrong)*: se está incorrecta;
- **X** *(ineXact)*: se contém mais ou menos informação do que a necessária;
- **U** *(Unsupported)*: se, apesar da resposta estar certa, o texto fornecido como suporte não contiver a resposta ou o texto não corresponder ao documento fornecido;
- **Z** *(Unknown)*: a resposta não foi avaliada;

As métricas utilizadas para medir e avaliar o desempenho dos sistemas no CLEF são [2]:

- **precisão**, a principal métrica utilizada, sobre a qual é feita a ordenação dos sistemas avaliados. Calcula a percentagem de perguntas respondidas correctamente, em relação ao número total de perguntas e é dada pela fórmula:

$$precisao = \sum_{q=0}^{Q} \frac{SCORE(q)}{Q} \tag{5.1}$$

em que:

$$SCORE(q) = \begin{cases} 1, se\ q\ avaliada\ como\ R, \\ 0, c.c. \end{cases}$$

- **K1**, obtida através da fórmula:

$$K1(sistema) = \sum_{r \in respostas(sistema)} \frac{score(r) \cdot eval(r)}{Q} \tag{5.2}$$

sendo que

$K1(sistema) \in R \cap K1(sistema) \in [-1, 1]$

e em que $score(r)$ é a confiança atribuída pelo sistema à pergunta r e $eval(r)$ depende da apreciação feita pelo juíz humano que avalia a resposta.

$$eval(r) = \begin{cases} 1, se\ r\ avaliada\ como\ correcta, \\ -1, c.c. \end{cases}$$

- **Confidence Weighted Score (CWS)**, que avalia as respostas por ordem decrescente de confiança que o sistema deposita nelas. Premeia os sistemas com respostas correctas nas primeiras posições da lista ordenada, de acordo com a seguinte fórmula:

[2] Definidas no documento *QA@CLEF07_Guidelines-for-Participants*, disponível em `http://clef-qa.itc.it/2007/guidelines.html`.

$$CWS = \frac{1}{Q}\sum_{i=1}^{Q}\frac{\#\ respostas\ correctas\ nas\ primeiras\ i\ posicoes}{i} \tag{5.3}$$

As métricas K1 e CWS medem a capacidade do sistema em avaliar as suas próprias respostas. O sistema QA@L^{2}F não possui a funcionalidade de se auto-avaliar, pelo que as referidas métricas não serão discutidas neste documento.

Em vez destas, e apesar desta não ter sido sujeita a avaliação no CLEF, a presente secção aborda a **cobertura** como métrica de desempenho. Utiliza a seguinte fórmula:

$$cobertura = \frac{TotalPerguntasRespondidas}{Q} \tag{5.4}$$

em que: $Q = 200$.

Nesta métrica, é calculada a percentagem de respostas para as quais o sistema tem, efectivamente, *scripts* de extracção de resposta em relação ao número total de perguntas. Para aquelas perguntas que não têm correspondência com um *script* de extracção de resposta, o sistema devolve NIL como resposta de omissão.

5.2.2 Resultados Obtidos

Nesta secção apresentam-se os resultados do sistema em cada uma das duas submissões efectuadas. Indicam-se e discutem-se os valores obtidos na sua avaliação. Apresentam-se, também, os gráficos comparativos dos resultados obtidos pelos sistemas portugueses no CLEF 2007.

Primeira Submissão

Os resultados obtidos pelo sistema na primeira submissão encontram-se na tabela 5.2.

	Nº de respostas	**% de respostas**	**Nº de respostas**	**% de respostas**
Right	22	**11.00**	17	14.91
Wrong	171	85.50	90	78.95
ineXact	5	2.50	5	4.39
Unsupported	2	1.00	2	1.75
Total	200	100.00	114	100.00

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

Na primeira submissão, o QA@L^{2}F teve uma precisão de 11%, com 22 correctas no total das 200 perguntas efectuadas ao sistema. Por outro lado, se tivermos em conta que a cobertura é de 57% (com

114 perguntas em 200 que o sistema tentou efectivamente responder), o número de respostas correctas é de 17, fazendo aumentar a sua precisão em 3,91 pontos percentuais, para 14.91%.

A tabela 5.3 apresenta os resultados da avaliação do sistema para cada uma das categorias das 200 perguntas efectuadas na primeira submissão. Após a análise da tabela verifica-se que o sistema obteve a maior precisão nas perguntas da categoria *Definition* (acertando 45.16%), sendo também nesta categoria que o sistema obteve as 5 respostas avaliadas como inexactas. Não acertou, porém, em nenhuma das 19 perguntas da categoria *List* (tendo uma precisão de 0%) pertencentes ao conjunto das 200 perguntas efectuadas. Das 151 perguntas devolvidas como NIL, por não ter encontrado resposta no *corpus* ou por não ter os respectivos *scripts* de extracção de resposta, o sistema acertou 9.

	Factoid	***Definition***	***List***	***Temp. Restricted***	***NIL***
Right	8	14	0	1	9
Wrong	150	11	10	18	142
ineXact	0	5	0	0	0
Unsupported	1	1	0	0	0
Total	159	31	10	19	151
Precisão	5.03%	45.16%	0.00%	5.26%	5.96%

Tabela 5.3: Resultados detalhados obtidos pelo QA@L^{2}F na 1ª submissão.

Das 10 perguntas da categoria *List* efectuadas, o sistema tentou responder a 7. A sua baixa *performance* deveu-se sobretudo ao facto de não conseguir distiguir este categoria de perguntas. Em nenhuma das 7 perguntas o sistema utilizou a abordagem implementada para esse efeito: a de reordenação de formulações linguísticas.

Segunda Submissão

Na tabela 5.4 são mostrados os resultados obtidos pelo sistema na segunda submissão. Esta submissão diferiu da primeira apenas num *script* de acesso e extracção de resposta na Wikipedia: o padrão para extracção de respostas foi alterado na estratégia de reordenação de formulações linguísticas.

	Nº de respostas	**% de respostas**	**Nº de respostas**	**% de respostas**
Right	26	**13.00**	21	18.42
Wrong	168	84.00	87	76.32
ineXact	4	2.00	4	3.51
Unsupported	2	1.00	2	1.75
Total	200	100.00	114	100.00

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

Na segunda submissão, o QA@L^{2}F teve uma precisão de 13%, com 26 correctas no total das 200

perguntas efectuadas ao sistema.

Novamente, a cobertura foi de 57%. Em 114 perguntas respondidas, 21 estavam certas, correspondendo a uma precisão de 18.42%.

O sistema reduziu o seu número de respostas inexactas para 4, porém estas 4 respostas mantiveram a mesma avaliação nas duas submissões. Para se perceberem estes resultados, analise-se o conjunto de perguntas, respectivas respostas e o pedaço de *corpus* de onde foram extraídas:

1: **Quem é Boaventura Kloppenburg?** bispo

Boaventura Kloppenburg, bispo de Novo Hamburgo (RS), disse...

2: **Quem foi Henrik Ibsen?** dramaturgo

Estou falando no Ibsen original, norueguês, Henrik Ibsen, dramaturgo que escreveu Peer Gynt .

3: **Quem é James Baker?** ex-secretário

O ex-secretário norte-americano de Estado James Baker recentemente previu

4: **Quem é George Vassiliou?** presidente de Chipre

George Vassiliou, presidente de Chipre entre 88 e 93, lançou, por sua vez, ■

Todas as perguntas utilizaram a estratégia *brute force* com pós-processamento de língua natural. Esta estratégia, aplicada a este tipo de perguntas, recolhe a entidade mencionada do tipo TITLE mais frequente de um conjunto de 10 parágrafos do *corpus*. Em todos os casos, a pergunta exigia mais do que a informação recolhida. Esta classificação coloca, no entanto, algumas dúvidas: será mesmo essencial, por exemplo, dizer que Henrik Ibsen foi o dramaturgo que escreveu Peer Gyint, não bastando a informação de que é um dramaturgo?

A tabela 5.5 apresenta detalhadamente os resultados da avaliação do sistema na sua segunda submissão, para cada uma das categorias de pergunta. Verifica-se, novamente, que o sistema obtém os melhores resultados para as perguntas da categoria *Definition*, tendo acertado em mais duas perguntas desta categoria em comparação com a primeira submissão. A precisão para esta categoria de perguntas aumentou 12.9% para 58.06%. O número de respostas NIL diminuiu em uma unidade, e destas, o número de respostas correctas manteve-se em 9. Os restantes resultados mantiveram-se iguais aos da primeira submissão.

O total de respostas não NIL correctas foi 17, das quais 3 foram respondidas pela estratégia de emparelhamento de padrões linguísticos e as restantes 14 pela estratégia de reordenação de formulações linguísticas. Verificou-se, também, que 2 utilizaram o mecanismo de relaxamento de restrições.

Quer na primeira submissão, quer na segunda, o sistema obteve resultados satisfatórios nas perguntas da categoria *Definition*. A taxa de respostas erradas foi inferior a 50% em cada um dos casos.

	Factoid	*Definition*	*List*	*Temp. Restricted*	*NIL*
Right	8	18	0	1	9
Wrong	150	8	10	18	141
ineXact	0	4	0	0	0
Unsupported	1	1	0	0	0
Total	159	31	10	19	150
Precisão	5.03%	58.06%	0.00%	5.26%	6.00%

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

De facto, verificou-se que, de entre as quatro estratégias existentes para extracção da resposta, o sistema utilizou, com sucesso, 3 delas:

- as perguntas "O que é a FIDE?" e "O que é a TVI?" foram respondidas acertadamente utilizando a abordagem de emparelhamento de padrões linguísticos;
- a pergunta "O que é um barrete frígio?" utilizou acertadamente a estratégia de reordenação de formulações linguísticas, assim como a pergunta "Mencione uma escritora sarda." que, apesar de não conter o parágrafo de suporte, está correcta;
- a pergunta "Quem é James Baker?" fez uso da abordagem *brute force* com pós-processamento linguístico. Apesar de inexacta, por estar incompleta (a resposta pretendida seria "ex-secretário norte-americano de Estado"), está certa.

Como conclusão desta segunda submissão e em relação às estratégias seguidas, observou-se que, do total das 200 perguntas:

- 17 utilizaram o emparelhamento de padrões linguísticos;
- 22 utilizaram a reordenação de formulações linguísticas;
- o emparelhamento de entidades mencionadas não deu resultados; e
- o *brute force* com pós-processamento linguístico conduziu a respostas inexactas.

Gráficos Comparativos

O gráfico 5.1 apresenta os resultados detalhados da avaliação das respostas para cada um dos sistemas concorrentes ao CLEF em 2007. [3]

[3]Foram considerados apenas os resultados para a melhor submissão dos sistemas.

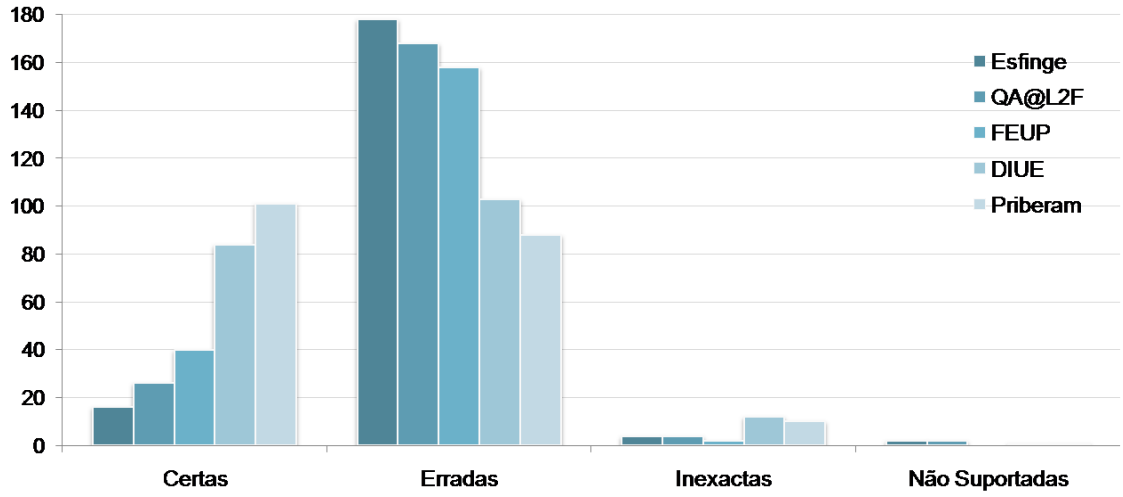

Figura 5.1: Gráfico comparativo das respostas dos sistemas portugueses no CLEF 2007.

O gráfico 5.2 apresenta os resultados comparativos da precisão dos sistemas portugueses participantes no CLEF em 2007, por ordem crescente desta métrica de avaliação.

Dos 5 sistemas portugueses avaliados, o da Priberam foi o que obteve melhores resultados, com uma precisão de 50,5%. Em segundo lugar, o sistema da Universidade de Évora com 42%, seguido do RAPOSA (o sistema de QA da Universidade do Porto) com 20%. O QA@L^{2}F posicionou-se em quarto lugar (13%) e em quinto lugar o Esfinge, com 8%.

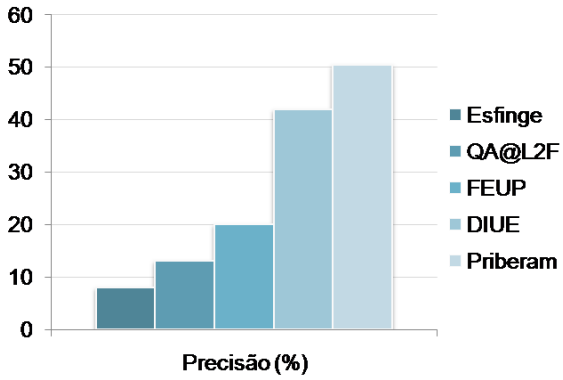

Figura 5.2: Gráfico comparativos da precisão dos sistemas portugueses no CLEF 2007.

5.3 Segunda Avaliação

A segunda avaliação teve como objectivo analisar a nova *performance* do sistema, relativamente a um novo conjunto de teste, também de 200 perguntas [4] de características semelhantes às utlizadas no CLEF

[4]Foi usado o conjunto de teste utilizado na edição do ano de 2004 do fórum CLEF.

em 2007, porém sem anáforas, elipses e perguntas da categoria *List*.

Esta avaliação foi efectuada depois da introdução de umas alterações ao sistema, em relação à quantidade de *scripts* existentes (que aumentou, permitindo responder a um maior número de perguntas, nomeadamente dos tipos `Quando...?` e `Quanto...?`), e de umas pequenas melhorias nos *scripts* implementados.

Resultados Obtidos

As respostas foram avaliadas apenas como correctas (C), incorrectas (I) ou inexactas (X), sendo a precisão, tal como definida pela equação 5.1, a métrica de desempenho utilizada. O conjunto de respostas certas possíveis foi fornecido com as 200 perguntas de teste. Os resultados estão na tabela 5.6.

	Nº de respostas	**% de respostas**
Correctas	29	**14.50**
Incorrectas	161	80.50
ineXactas	10	5.00
Total	200	100.00

Tabela 5.6: Resultados obtidos pelo QA@L^{2}F na 2ª avaliação.

A precisão do sistema nesta segunda avaliação mediu-se em 14.50 pontos percentuais com 29 respostas correctas em 200. No entanto, destas, apenas 3 corresponderam a respostas NIL. Ou seja, o sistema encontrou a resposta pretendida, correcta e exacta para 26 perguntas, quer no *corpus*, quer na Wikipedia.

Os resultados detalhados da avaliação, em função de cada uma das categorias de pergunta existentes, estão apresentados na tabela 5.7.

Tal como na avaliação no CLEF 2007, é nas perguntas da categoria *Definition* que o sistema obtém os melhores resultados: 13 respostas certas em 30 efectuadas desta categoria. Por outro lado, a precisão nas perguntas da categoria *Factoid* é demasiado baixa, de apenas 9.40%, com 16 respostas certas em 170.

	Factoid	***Definition***
Correctas	16	13
Incorrectas	149	12
ineXactas	5	5
Total	170	30
Precisão	9.41%	43.30%

Tabela 5.7: Resultados detalhados obtidos pelo QA@L^{2}F na 2ª avaliação.

Depois da recolha destes resultados do sistema, devem ser discutidos três pontos:

- O número de respostas inexactas corresponde a 5.00% do total. A decisão de avaliar uma resposta como inexacta é, novamente, subjectiva. Considerem-se as seguintes três perguntas, respectivas respostas dadas pelo sistema e as frases do *corpus* que lhes servem de suporte:

  1: **Quem é Jorge Amado?** escritor

  Qual o endereço para correspondência do escritor Jorge Amado?...

  2: **Quem é Win Dulsenberg?** presidente

  O presidente do banco central holandês, Win Dulsenberg...

  3: **Qual o nome da mulher de Kurt Cobain?** Courtney

  Cobain tinha 27 anos e uma filha de dois, Frances. Sua mulher, Courtney Love,... ■

  Se por um lado se pode tomar a profissão de escritor como uma correcta definição de Jorge Amado, tal não fará sentido no caso de Win Dulsenberg: o termo presidente é demasiado geral para constituir a definição única de uma pessoa. Por outro lado, como se encontra patente na frase retirada do *corpus*, o nome exacto e comummente utilizado da mulher de Kurt Cobain é Courtney Love. As respostas às perguntas 2 e 3 foram avaliadas como inexactas, apesar de correctas.

- nas perguntas do tipo `Onde...?`, a exactidão da resposta foi um elemento decisivo. Considerem-se as seguintes perguntas:

  1: **Onde é Izhesvk?** Moscovo

  ...em Izhesvk, a 960 quilómetros a leste de Moscovo.

  2: **Onde fica a Esfinge de Gizé?** Cairo

  ...defendem que a Esfinge de Gizé, perto do Cairo,... ■

  Ambas as respostas dadas conseguem ajudar o utilizador a localizar quer Izhesvk, quer a Esfinge de Gizé, não correspondendo, no entanto, às respostas correctas. Apesar da relação "perto de" ser subjectiva e inquantificável, no entanto, geograficamente, uma distância de 960 quilómetros justifica a inexactidão da resposta. Assim, as respostas anteriores foram avaliadas como inexactas.

- O sistema baseia-se profundamente na identificação e recolha de entidades mencionadas, que podem ser decisivas na resposta a uma pergunta. A seguinte, por exemplo, teve uma resposta incorrecta que se justifica com a não classificação da palavra *hacker* como entidade mencionada do tipo TITLE. De outra forma, a relação entre a pessoa Kevin Mitnick e o seu título de *hacker* seria capturada através da estratégia de emparelhamento de padrões linguísticos ou, caso tal não acontecesse, na estratégia *brute-force* com pós-processamento de língua natural:

1: **Quem é Kevin Mitnick?** segurança

A audácia do *hacker* Kevin Mitnick,... ■

A dependência em demasia na identificação e recolha de entidades mencionadas é um ponto menos positivo na avaliação do sistema. Este não lida bem com perguntas cuja resposta não seja uma entidade mencionada, bem como com perguntas para as quais a resposta seja uma entidade mencionada composta com outra informação.

Apesar da maior ou menor flexibilidade dos critérios de avaliação para classificar as respostas como inexactas e do facto destas serem uma fonte de resultados menos bons do sistema, há que notar que o sistema teve 80.50% de respostas incorrectas. Nestas contam-se aquelas que ainda não são tratadas pelo sistema, bem como aquelas a que tenta responder, mas cuja resposta está errada. A tabela 5.8 apresenta o resumo das respostas erradas para o conjunto das 200 perguntas:

	Nº de respostas	**% de respostas**
Não tratadas	20	10.00%
Resposta Errada	141	70.50%
Total	161	80.50%

Tabela 5.8: Resumo das respostas erradas na 2ª avaliação.

Seguidamente analisar-se-ão um conjunto de situações que levaram a estes valores de respostas incorrectas.

Em primeiro lugar, a informação recolhida na fase de análise e interpretação da pergunta é demasiado escassa, chegando por vezes a ser nula. Em 7% dos casos (14 perguntas) a *frame* construída foi a seguinte:

Frame

```
SCRIPT script-what.pl
TARGET
ENTIDADES
```
■

Os *scripts* implementados não estão preparados para receber este tipo de *frames*, provocando respostas erradas em todas elas.

Por outro lado, e apesar de unificar a informação recebida e extraída da pergunta com o *corpus*, essa abordagem não é suficiente. Considere-se a pergunta, resposta dada e parágrafo de onde foi recolhida:

1: **Em que estado do Brasil fica Campo Grande?** Flórida

Gadig está de férias em Santos e visitou a vítima, que mora no bairro de Campo Grande. Ele é representante no Brasil do Museu Estadual da Flórida (EUA), que cataloga ataques de tubarões no mundo. Na região da Baixada Santista, temos registrados quatro ataques de tubarões desde 1976, incluindo uma morte , disse o biólogo. ■

O sistema identificou Flórida como estado nesta frase e retornou-a como resposta. No entanto, a resposta pretendida era mais restritiva: pretendia-se um estado do Brasil, ao invés de apenas um estado. O sistema não tem capacidade de reconhecer e tratar este tipo de restrições.

Noutro exemplo, considere-se a seguinte resposta dada pelo sistema:

1: **Que grupo matou Aldo Moro?** Maccari

Um tribunal de Roma pôs ontem em liberdade Germano Maccari, suspeito de ser um dos assassinos do ex-primeiro-ministro italiano Aldo Moro, morto em 1978 pelas Brigadas Vermelhas. O tribunal de revisão das detenções preventivas rejeitou um pedido do Ministério Público para manter Maccari na prisão por mais seis meses. Maccari, de 41 anos, foi preso em 14 de Outubro de 1993 pela polícia anti-terrorista e foi apontado na mesma época por uma antiga militante das Brigadas, Adriana Faranda, como o autor do golpe de misericórdia que matou Moro. ■

Tendo conseguido identificar com sucesso o parágrafo do *corpus* contendo a informação necessária, o sistema falhou na devolução da resposta final. De facto, apesar da pergunta pedir um grupo (e, portanto, uma entidade mencionda do tipo ORG), o sistema devolveu como resposta a entidade do tipo PROPER mais frequente, o que se justifica com as seguintes razões[5]:

- a fase de análise e interpretação da pergunta encaminhou a informação para o `script-what.pl`, ao invés de a enviar para o `script-who-org.pl` que se destina a recolher entidades mencionadas do tipo ORG;
- o `script-what.pl` encaminhou a informação para o `script-what-chaos.pl`, que aplica a estratégia *brute force* com pós-processamento de língua natural e que recolhe a entidade mencionada mais frequente do tipo PROPER;
- utilizando a abordagem de recolher a entidade mencionada do tipo pretendido (ORG), o sistema não retornaria a resposta certa já que Brigadas Vermelhas não são identificadas como entidade

[5]De referir que o nome dos *scripts* aqui apresentados não correspondem na sua totalidade aos *scripts* descritos na secção 4.4 deste documento. Tal deve-se às alterações efectuadas no sistema para esta segunda avaliação.

mencionada deste tipo.

A pergunta seguinte teve também uma resposta incorrecta, mas desta vez por outros motivos:

1: **Quantos submarinos tem a marinha portuguesa?** dois

O [Barracuda], construído em França há 27 anos, é um dos três submarinos da marinha portuguesa – os outros são o [Albacora] e o [Delfim] –, cuja missões principais são o reconhecimento e a vigilância da costa portuguesa. Apenas dois dos submersíveis estão, em simultâneo, operacionais, dado se revezarem periodicamente para grandes revisões. ■

Novamente, a resposta correcta encontra-se no parágrafo dado como suporte, sendo que a resposta dada é também do tipo pretendido: NUM. Porém, na existência de duas palavras com a mesma frequência, o sistema recolheu a última como resposta final. Foi, no entanto, a escolha errada.

5.4 Sumário

Este capítulo apresentou os resultados do sistema em cada um dos momentos em que foi avaliado, bem como as medidas e métricas de desempenho utilizadas.

A participação no fórum CLEF resultou nos valores de 13% de respostas correctas, em 200 perguntas efectuadas. Já na segunda avaliação, feita no laboratório, e tendo sofrido algumas alterações, o sistema conseguiu resultados de precisão na ordem dos 14.50%.

Foram discutidos os resultados, bem como apresentadas algumas razões para os valores obtidos.

6 Conclusão

The important thing is not to stop questioning.

– Albert Einstein

O QA@L^{2}F é um sistema de QA desenvolvido no L^{2}F que utiliza um conjunto de quatro estratégias diferentes por forma a responder às perguntas que lhe são submetidas. Faz uso de uma análise linguística profunda providenciada por uma cadeia de PLN que lhe fornece informação morfológica, sintáctica e semântica, nomeadamente entidades mencionadas e dependências, quer acerca da pergunta, quer sobre o *corpus*.

Apesar dos resultados da sua avaliação não serem elevados (tendo atingindo 14.50% de precisão), é de referir que o QA@L^{2}F constitui apenas os primeiros passos de um sistema que irá crescer e desenvolver-se, nesta área bastante *sui generis* que é a dos sistemas de *question-answering*.

O sistema permitiu testar um conjunto de diferentes estratégias para extracção das resposta final. Apesar de implementadas em largura, e não em profundidade, uma das conclusões a retirar é a de que existe um conjunto de estratégias que se adequam a determinado tipo de pergunta de entrada, bem como *corpus* que permite responder mais efectivamente a determinadas perguntas. Por exemplo, as perguntas do tipo *definition* foram resolvidas utilizando a estratégias baseadas em padrões, que façam a procura na Wikipedia.

Relativamente ao pré-processamento do *corpus*, concluiu-se que esta é uma etapa em que se deve investir. Permitindo a estruturação da informação *a priori*, através da criação de dependências entre conceitos, e a geração de uma base de conhecimento (como é o caso da tabela de factos no QA@L^{2}F), esta fase permite que, após a submissão da pergunta pelo utilizador, a informação seja mais facil e rapidamente acedida.

A detecção de entidades mencionadas é de importância determinante para o QA@L^{2}F. Por um lado, como aspecto positivo, grande parte das perguntas efectuadas a um sistema deste tipo esperam como resposta uma entidade mencionada; por outro lado, como aspecto negativo, a dependência naquelas origina muitas vezes inexactidão nas respostas.

Finalmente, confirmou-se que, por vezes, as estratégias mais simples são as que obtêm os melhores resultados, como é o caso da *brute force* com pós-processamento de língua natural. Apesar de ser

preterida a outras estratégias vistas inicialmente como mais apropriadas para a resolução de determinadas perguntas (a estratégia *brute force* era muita vezes denominada de estratégia do caos, sendo que os *scripts* que a implementam têm a palavra *chaos*), foi com o recurso a esta estratégia que muitas perguntas tiveram resposta certa ou inexacta.

6.1 Contribuições

O sistema desenvolvido permitiu testar um conjunto de abordagens diferentes na tarefa de QA, fazendo uso da cadeia de processamento de língua natural que, à excepção do último módulo (XIP), foi inteiramente desenvolvida no laboratório.

É um sistema funcional que lida com perguntas de domínio aberto. A arquitectura de base desenvolvida para o sistema permite a implementação de novas estratégias para extracção da resposta e a melhoria das já existentes, bem como o desenvolvimento de cada módulo de forma independente, desde que se definam convenientemente as interfaces de comunicação entre eles.

A participação no fórum de avaliação CLEF, possibilitou a divulgação do L^2F junto da comunidade científica nacional e internacional, interessada na investigação da língua natural e seu processamento computacional, desta vez na área dos sistemas de QA.

6.2 Trabalho Futuro

Tendo esta tese como título "QA@L^2F: primeiros passos", não será de espantar que a secção relativa ao trabalho futuro, e que aqui inicia, ocupe um espaço razoável neste documento.

Algumas ideias nascidas para a concepção deste sistema foram de facto implementadas. No entanto, com o avançar do tempo e do trabalho e com a maior familiaridade que com ele era gerada, muitas outras surgiram e que não passaram da fase de incubação.

6.2.1 Extensões

Esta secção descreve algumas funcionalidades que, apesar de implementadas no sistema, poderiam ser melhoradas, por forma a aumentar, também, a qualidade do sistema.

Perguntas Tratadas

Na sua situação actual, o leque de perguntas que o sistema trata ainda é muito reduzido, limitando-se às perguntas que iniciam com as habituais formas: `Quem...?`, `O que...?`, `Onde...?` ou `Quando...?`.

Perguntas que fujam a este modelo, como, por exemplo, "Depois de 1995, que clubes ganharam a Taça de Portugal?", ainda não são processadas, o mesmo acontecendo para as que utilizam recursos estilísticos como a anáfora e a elipse.

Mecanismo de *Frames*

O sistema não tem implementadas funcionalidades que lhe permitam recolher e processar toda a informação contida na pergunta. Palavras e relações entre elas muitas vezes determinantes nas perguntas e das quais pode depender o seu sentido semântico, não têm significado prático, não sendo tratadas devidamente. O sistema limita-se a tentar fazer o emparelhamento entre elas e o *corpus*. Por exemplo, as duas perguntas dadas em seguida como exemplo são tratadas de forma igual pelo sistema.

Exemplo 10:

Pergunta 1: De quem é filha Maria de Medeiros?

Pergunta 2: Quem é a filha de Maria de Medeiros? ■

Ambas esperam uma entidade mencionada do tipo PEOPLE como resposta, e contêm as entidades RELATIVE filha e PEOPLE Maria de Medeiros. Têm, no entanto, semânticas completamente distintas, sendo as suas respostas diferentes.

Outras palavras e relações não convenientemente tratadas são, por exemplo, marcas temporais ("antes de", "depois de", "até"), marcas de sequência ("primeiro", "último"), negações ("não", "nunca"), grandezas ("maior", "mais pequeno", "mais novo").

Desta forma, seria útil dotar o mecanismo de *frames* utilizado na etapa de análise e interpretação da pergunta de uma maior expressividade, bem como fazer um processamento mais profundo das relações existentes no *corpus*, que lhe permita tratar convenientemente estes casos.

Estratégia de Reordenação de Formulações Linguísticas

Estender a estratégia de reordenação de formulações linguísticas, a outras perguntas que não apenas do tipo *definition* e *list*. A ideia seria a de transformar as perguntas nos padrões de resposta que lhes correspondem directamente, efectuando depois uma procura no *corpus* com esses padrões, sem recorrer a outro tipo de processamento.

O exemplo seguinte mostra uma pergunta e alguns padrões que lhe correspondem, sendo $R o local onde a resposta poderá aparecer.

Exemplo 11:

Pergunta: Em que data é que os Estados Unidos invadiram o Haiti?

Padrão 1: os Estados Unidos invadiram o Haiti em $R

Padrão 2: os Estados Unidos, em $R, invadiram o Haiti

Padrão 3: os Estados Unidos invadiram o Haiti no ano $R ■

Escolha da Resposta Final

Seria interessante o módulo de extracção da resposta final, na abordagem de emparelhamento de entidades mencionadas, basear-se noutro método que não a pura frequência de conceitos. Dando pesos às respostas obtidas, por exemplo, de acordo com a maior ou menor proximidade a que se encontram das entidades mencionadas seria uma estratégia a considerar.

6.2.2 Novas Funcionalidades

Esta secção apresenta algumas funcionalidades não implementadas no sistema, cujo desenvolvimento poderia contribuir para aumentar a sua eficiência.

A Resposta na Sintaxe

A sintaxe pode também ajudar na descoberta da resposta a determinadas perguntas. Tomando em consideração um conjunto de perguntas que podem ser efectuadas ao sistema (como, por exemplo, as utilizadas no fórum CLEF), verifica-se que as respostas a grande parte delas se podem encontrar recorrendo apenas a uma análise morfo-sintáctica do *corpus*.

Por exemplo, às perguntas do tipo `Onde...?`, o sistema poderá encontrar a resposta num complemento circunstancial de lugar associado a uma forma verbal do verbo principal presente na pergunta; às do tipo `Quando...?`, num complemento circunstancial de tempo; às do tipo `Com quem...?`, num complemento circunstancial de companhia; e assim por diante.

Módulo de Validação da Resposta

A inclusão de um módulo de validação da resposta, no final da cadeia de processamento do QA@L^{2}F, teria a função de retroalimentar o sistema. Recebendo como entrada a pegunta efectuada, a resposta encontrada pelo sistema e a frase do *corpus* que a suporta, a sua função seria garantir que a resposta dada se adequa à pergunta e se se encontra na frase em questão. Caso tal aconteça, a resposta pode

ser devolvida ao utilizador como final; caso contrário, o sistema deve procurar uma nova resposta à pergunta submetida.

De forma a ilustrar-se a funcionalidade deste módulo, considere-se o exemplo seguinte:

Exemplo 12:

Pergunta: De que cor é a neve?

Resposta: Branca

Frase: A Walt Disney recebeu encomendas na ordem das 27 milhões de unidade para o lançamento no mercado de vídeo dos EUA do seu filme Branca de Neve, marcado para 25 de Outubro. ■

De facto, Branca pode ser uma resposta correcta à pergunta formulada. No entanto, a frase fornecida não é válida como suporte. As semânticas associadas à pergunta e à frase são completamente distintas, originando a não exactidão da resposta. Assim, o módulo de validação de respostas deveria informar o sistema de QA acerca da inviabilidade da frase de suporte, de modo a que o sistema faça uma nova tentativa na procura da resposta certa e respectiva frase.

Devido à complexidade em implementar um módulo com estas características, este desafio corresponde a uma tarefa avaliada no fórum CLEF.

Subdivisão em Domínios

A subdivisão em domínios implicaria um maior grau de processamento do *corpus*, nomeadamente na identificação e recolha de padrões e relações de dependência entre conceitos. A ideia seria a de dividir e estruturar a base de dados em domínios, como literatura, música, cinema, política. A fase de pré-processamento teria a responsabilidade de a popular de acordo com estas subdivisões, respeitando os domínios existentes e criando as associações entre as entradas nas tabelas, por forma a corresponderem às relações descritas no *corpus*. Esta aproximação traria como vantagens:

- selecção, de acordo com o domínio da pergunta, pelo tipo de informação pretendida;
- redução da quantidade de informação na qual é necessário fazer a procura, focando-se apenas no domínio que interessa;
- possibilidade de melhoria de cada domínio de forma independente.

Neste sentido, com a base de dados subdividida em domínios, existiria a possibilidade de integração com o trabalho desenvolvido por (Guimarães, 2007), uma interface em língua natural para base de dados, no domínio específico do cinema, que conta com uma estrutura de base de dados para cinema e forma de lhe aceder bem definidas.

Procura na *Internet*

Além das estratégias já implementadas para a procura e extracção da resposta final, implementar uma que utilizasse a *Internet* como fonte de informação.

A estratégia implicaria enviar toda a informação disponível (sem qualquer tipo de processamento) para um motor de busca, como o Google[1]. Seguidamente, recolher o conjunto dos n documentos melhor classificados e fazer um processamento de língua natural que permitisse recolher entidades mencionadas. Finalmente, analisar o resultado do processamento e devolver como resposta final a entidade mencionada do tipo esperado mais frequente. Esta estratégia é semelhante à *brute force* com pós-processamento de língua natural; no entanto, em vez de recorrer ao *corpus* jornalístico armazenado em base de dados, utilizaria a *Internet* como recurso.

[1] `http://www.google.com`

Bibliografia

Amaral, C., Figueira, H., Martins, A., Mendes, A., Mendes, P., & Pinto, C. (2005). Priberam's question answering system for Portuguese. *Working Notes for the CLEF 2005 Workshop*, 21–23.

Aït-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.

Bick, E. (2000). *The Parsing System PALAVRAS - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Bixler, D., Moldovan, D., & Fowler., A. (2005). Using Knowledge Extraction and Maintenance Techniques to Enhance Analytical Performance. In *Proceedings of the 2005 International Conference on Intelligence Analysis.* Washington D.C.

Bouma, G., Fahmi, I., Mur, J., Noord, G. van, Plas, L. van der, & Tiedmann, J. (2006). The University of Groningen at QA@CLEF 2006 Using Syntatic Knowledge for QA. *Working Notes for the CLEF 2006 Workshop*.

Bouma, G., Noord, G. van, & Malouf, R. (2000). *Alpino: wide-coverage computational analysis of Dutch.*

Brill, E. (2003). Processing Natural Language without Natural Language Processing. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics* (p. 360-9).

Burch, R. (2007). Charles Sanders Peirce. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.*

Buscaldi, D., Gomez, J. M., Rosso, P., & Sanchis, E. (2006). The UPV at QA@CLEF 2006. *Working Notes for the CLEF 2006 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*.

Costa, L. (2004). First evaluation of Esfinge - a question answering system for Portuguese. *Working Notes for the CLEF 2004 Workshop*.

Costa, L. (2005). 20th century esfinge (Sphinx) solving the riddles at CLEF 2005. *Working Notes for the CLEF 2005 Workshop*.

Costa, L. (2006). Esfinge - a modular question answering system for portuguese. *Working Notes for the CLEF 2006 Workshop*.

Filho, P. P. B., Uzêda, V. R. de, Pardo, T. A. S., & Graças Volpe Nunes, M. das. (2006). Using a Text Summarization System for Monolingual Question Answering. *Working Notes for the CLEF 2006 Workshop*.

Guimarães, A. R. (2007). *JáTeDigo — Uma interface em língua natural para uma base de dados de cinema*. Unpublished master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Harabagiu, S., Moldovan, D., Christine Clark, M. B., Hickl, A., & Wang, P. (2005). Employing Two Question Answering Systems in TREC-2005. In *Proceedings of the Fourteenth Text REtrieval Conference.*

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Juárez-Gonzalez, A., Téllez-Valero, A., Denicia-Carral, C., Gómez, M. M. y, & Villaseñor-Pineda, L. (2006). INAOE at CLEF 2006: Experiments in Spanish Question Answering. *Working Notes for the CLEF 2006 Workshop*.

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers.

Laurent, D., Séguéla, P., & Nègre, S. (2006). Cross Lingual Question Answer using QRISTAL for CLEF 2006. *Working Notes for the CLEF 2006 Workshop*.

Loureiro, J. (2007). *Reconhecimento de Entidades Mencionadas (Tempo, Valor, Relações de Parentesco e Obra) e Normalização de Expressões Temporais*. Unpublished master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Mamede, N. J. (2007). *A Cadeia de Processamento XIP em Maio de 2007.*

Medeiros, J. C. (1995). *Análise morfológica e correcção ortográfica do português*. Unpublished master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Mendes, A., Coheur, L., Mamede, N. J., Romão, L., Loureiro, J., Ribeiro, R., et al. (2007). QA@L^{2}F@QA@CLEF. *Working Notes for the CLEF 2007 Workshop*.

Moldovan, D., Clark, C., Harabagiu, S., & Maiorano, S. (2003). COGEX: a logic prover for question answering. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 87–93). Morristown, NJ, USA: Association for Computational Linguistics.

Neumann, G., & Piskorski, J. (2002). A Shallow Text Processing Core Engine. *Journal of Computational Intelligence*, *18*(3).

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001.*

Pardal, J. P., & Mamede, N. J. (2004, Novembro). Terms Spotting with Linguistics and Statistics. 298-304.

Pardo, T. A. S. (2006, Janeiro). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil* (Tech. Rep.). Núcleo Interinstitucional de Linguística Computacional.

Peters, C., et al. (Eds.). (2007). *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers.* Springer.

Quaresma, P., Quintano, L., Rodrigues, I., Saias, J., & Salgueiro, P. (2004). The University of Évora approach to QA@CLEF-2004. *Working Notes for the CLEF 2004 Workshop*.

Quaresma, P., & Rodrigues, I. (2005). A logic programming based approach to the QA@CLEF05 track. *Working Notes for the CLEF 2005 Workshop*.

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.

Rocha, P., & Santos, D. (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, 131–140.

Rodrigues, D. J. (2007). *Uma evolução no sistema ShRep: optimização, interface gráfica e integração de mais ferramentas*. Unpublished master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Romão, L. (2007). *Reconhecimento de Entidades Mencionadas em Lìngua Portuguesa: Locais, Pessoas, Organizações e Acontecimentos*. Unpublished master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Sacaleanu, B., & Neumann, G. (2006). DFKI-LT at the CLEF 2006 Multiple Language Question Answering Track. *Working Notes for the CLEF 2006 Workshop*.

Santos, D., & Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for Portuguese. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 450–457.

Sarmento, L. (2006a, 22-28 May). BACO - A large database of text and co-occurrences. In Nicoletta Calzolari and Khalid Choukri and Aldo Gangemi and Bente Maegaard and Joseph Mariani and Jan Odjik and Daniel Tapias (Ed.), *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (p. 1787-1790). Genova, Italia.

Sarmento, L. (2006b). Hunting answers with RAPOSA (FOX). *Working Notes for the CLEF 2006 Workshop*.

Sarmento, L. (2006c, JanuaryMarch-JanuaryJuly May). SIEMÊS - a named entity recognizer for Portuguese relying on similarity rules. In R. Vieira, P. Quaresma, Maria, N. Mamede, C. Oliveira, & M. C. Dias (Eds.), *Proc. of the 7th Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006* (pp. 90–99). Itatiaia, Rio de Janeiro, Brazil: Springer.

Voorhees, E. M. (2003). Overview of the TREC 2003 Question Answering Track. In L. Buckland & E.Voorhees (Eds.), *Proc. of TREC 2003, NIST, Gaithersburg, USA.*

Voorhees, E. M., & Buckland, L. P. (Eds.). (2005). *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005).*

WebsiteCLEF. (n.d.). *http://www.clef-campaign.org.* (Visitado em Setembro de 2007)

WebsiteEsfinge. (n.d.). *http://www.linguateca.pt/Esfinge.* (Visitado em Setembro de 2007)

WebsiteHAREM. (n.d.). *http://poloxldb.linguateca.pt/harem.php.* (Visitado em Setembro de 2007)

WebsiteLCC. (n.d.). *http://www.languagecomputer.com/.* (Visitado em Setembro de 2007)

WebsiteM-Cast. (n.d.). *http://www.m-cast.infovide.pl/portugues/index.html.* (Visitado em Setembro de 2007)

WebsiteNTCIR. (n.d.). *http://research.nii.ac.jp/ntcir/workshop.* (Visitado em Setembro de 2007)

WebsiteTREC. (n.d.). *http://trec.nist.gov.* (Visitado em Setembro de 2007)

WebsiteWordnet. (n.d.). *http://xwn.hlt.utdallas.edu.* (Visitado em Setembro de 2007)

I Apêndices

A Avaliação no CLEF 2007

Neste apêndice são apresentadas as 200 perguntas que constituíram o conjunto de teste do sistema. Para cada pergunta são apresentados o seu identificador, o identificador do grupo a que pertence, a sua categoria, se tem restrições temporais e o resultado da sua avaliação na primeira e segunda submissões:

ID	Grupo	Cat.	R.Temporal	Pergunta	I	II
0001	1800	F	Sim	Que país declarou a independência em 1291?	W	W
0002	1801	F	Sim	O que se passou a 4 de Julho de 1776?	W	W
0003	1802	L	Não	De quem trata o filme "2 filhos de Francisco"?	W	W
0004	1803	L	Não	Quais são as sete colinas de Roma?	W	W
0005	1803	F	Não	Qual é a mais pequena delas?	W	W
0006	1803	F	Não	Quem é o dono delas?	R	R
0007	1803	F	Não	Em qual delas fica a residência do presidente?	W	W
0008	1804	F	Não	De que estado foi governador Adhemar de Barros?	W	W
0009	1805	D	Não	Quem eram os almóadas?	W	R
0010	1806	F	Não	Quem fabrica os foguetões Ariane?	W	W
0011	1807	F	Não	Quantas pessoas morreram no atentado na estação de Bolonha?	W	W
0012	1808	D	Não	Quem é Boaventura Kloppenburg?	X	X
0013	1809	F	Não	Quando é que a baleia azul foi extinta?	R	R
0014	1810	F	Não	O que está escrito na bandeira do Brasil?	W	W
0015	1811	F	Não	Que árvore está na bandeira do Líbano?	W	W
0016	1812	F	Não	Por que estado é senador Barack Obama?	W	W
0017	1813	D	Não	O que é um barrete frígio?	R	R
0018	1814	F	Não	Quantos jogadores tem um time de basquete?	W	W
0019	1815	F	Não	Quem disse "Ao vencedor, as batatas!"?	W	W
0020	1816	F	Não	Que princesa portuguesa foi rainha da Dinamarca?	W	W
0021	1817	F	Não	Que inglês foi treinador do Porto?	W	W
0022	1818	F	Não	Que mar banha Braga?	R	R
0023	1819	D	Não	O que é a navegação de cabotagem?	R	R
0024	1820	F	Não	Caen é a capital de que departamento?	W	W
0025	1821	D	Não	O que é um caiaque?	R	R
0026	1822	F	Não	Como se chama Cayenne em português?	W	W
0027	1823	L	Não	Quem viajava no avião que se despenhou em Camarate?	W	W
0028	1824	F	Sim	Que cidade era capital do Brasil antes do governo ser transferido para Brasília?	W	W
0029	1825	F	Sim	Que rei morreu em 1718?	W	W
0030	1826	D	Não	O que eram cartas de alforria?	W	W
0031	1827	F	Não	Que cidade francesa foi residência dos Papas?	W	W
0032	1828	F	Não	Quando foi fundado o Nacional da Madeira?	W	W
0033	1828	F	Não	Onde?	W	W
0034	1829	F	Não	Qual o diâmetro de Ceres?	W	W
0035	1830	F	Não	Diga o nome de um chá branco.	W	W
0036	1831	F	Não	Em que distrito fica Chaves?	W	W

ID	Grupo	Cat.	R.Temporal	Pergunta	I	II
0037	1831	F	Não	Quantas freguesias tem o concelho?	W	W
0038	1831	L	Não	Que municípios ficam a oeste?	W	W
0039	1831	F	Sim	Quantos habitantes tinha em 2006?	W	W
0040	1832	L	Não	Quais são as cidades-estado da Alemanha?	W	W
0041	1833	F	Não	De onde é nativo o cisne branco?	W	W
0042	1834	F	Não	A quem pertence o Cisne Branco?	W	W
0043	1835	F	Não	Quando foi registado pela primeira vez o cometa Halley?	W	W
0044	1835	F	Não	Em que famosa obra aparece o cometa?	W	W
0045	1835	F	Não	Que batalha aparece nessa obra?	W	W
0046	1836	F	Sim	Quem venceu o Campeonato do Mundo de Contra-Relógio em 1995?	W	W
0047	1836	F	Não	Em que país se disputou a prova?	W	W
0048	1836	F	Não	Qual o tempo do vencedor?	W	W
0049	1836	F	Sim	E quem venceu em 1994?	W	W
0050	1837	F	Não	Com quem se casou a rainha Cristina?	R	R
0051	1837	F	Não	Que idade tinha ela quando morreu?	W	W
0052	1838	F	Não	Quem foi o 13º rei de Portugal?	W	W
0053	1839	D	Não	O que é a defesa siciliana?	R	R
0054	1840	F	Não	Que empresa editou a Diciopédia?	W	W
0055	1841	F	Sim	Que dinastia reinava em Portugal em 1500?	W	W
0056	1842	D	Não	O que são os DOM franceses?	W	W
0057	1842	F	Não	Quantos são eles?	W	W
0058	1842	F	Não	Qual deles fica na Europa?	R	R
0059	1843	F	Não	Em que zona de Espanha nasce o rio Douro?	W	W
0060	1844	F	Não	Onde fica o parque Eduardo VII?	W	W
0061	1845	F	Não	Quantos focos tem uma elipse?	W	W
0062	1846	F	Não	Quantos pontos vale um ensaio no râguebi?	W	W
0063	1847	D	Não	O que é o enxaimel?	R	R
0064	1848	F	Não	Diga uma escritora sarda.	U	U
0065	1849	F	Não	Como se chamava o sabre dos samurai?	W	W
0066	1850	D	Não	Quem foi Ésquilo?	W	R
0067	1851	F	Não	Em que constelação fica Vega?	W	W
0068	1851	F	Não	E Régulo?	W	W
0069	1852	F	Não	Qual a freguesia mais populosa de Évora?	W	W
0070	1853	F	Não	Que equipa é treinada por Fernando Santos?	W	W
0071	1854	F	Sim	Que equipa era treinada por Fernando Santos em 1994?	W	W
0072	1855	F	Não	Em que estado brasileiro fica Faro?	W	W
0073	1856	F	Não	Em que época é que o FC Porto ganhou a Liga dos Campeões?	W	W
0074	1857	F	Não	Que pássaro renascia das próprias cinzas?	W	W
0075	1858	F	Não	Que percentagem dos finlandeses fala sueco?	W	W
0076	1859	F	Não	Quando é que F.H. Cardoso tomou posse como presidente?	W	W
0077	1860	D	Não	O que é a FIDE?	R	R
0078	1861	F	Não	Que profundidade atinge a Fossa das Marianas?	W	W
0079	1861	F	Não	Onde fica?	W	W
0080	1861	F	Não	Quando se atingiu o seu fundo?	W	W
0081	1861	F	Não	Por que veículo?	W	W
0082	1862	F	Não	Quantas horas há de diferença entre Nova Iorque e São Francisco?	W	W
0083	1863	F	Não	Além da girafa, que outro animal pertence à família Giraffidae?	W	W
0084	1864	F	Não	Que espada usavam as legiões romanas?	W	W
0085	1865	F	Não	Quando caíu Granada?	W	W
0086	1866	D	Não	O que é a Granja do Torto?	R	R
0087	1867	F	Não	Qual é a latitude e longitude da Guarda?	W	W
0088	1868	F	Não	De que estado foi Hans Modrow primeiro-ministro?	W	W
0089	1869	D	Não	Quem foi Henrik Ibsen?	X	X

ID	Grupo	Cat.	R.Temporal	Pergunta	I	II
0090	1870	F	Não	Quem matou Philipe Henriot?	W	W
0091	1871	D	Não	O que é o ICBAS?	R	R
0092	1872	F	Não	Qual a população da Ilha de Moçambique?	W	W
0093	1873	F	Não	Qual o estado mais populoso da Índia?	W	W
0094	1874	D	Não	O que é o jagartee?	R	R
0095	1875	D	Não	Quem é James Baker?	X	X
0096	1876	L	Não	Como se chamam os filhos de Jimmy Carter?	W	W
0097	1877	F	Não	Quando se realizaram os Jogos Olímpicos de Munique?	W	W
0098	1878	F	Não	Que senador americano foi prisioneiro no Vietname?	W	W
0099	1879	F	Não	Em que dia ocorreu a batalha de La Lys?	W	W
0100	1880	F	Não	Em que ilha dos Açores fica a Lagoa?	W	W
0101	1881	D	Não	Quem é o Lampadinha?	W	W
0102	1882	F	Não	O lehendakari é presidente de quê?	W	W
0103	1883	F	Não	Quando é que foi assinada a Lei Áurea?	W	W
0104	1883	F	Não	Por quem?	W	W
0105	1884	F	Não	Que é o que Arafat usava na cabeça?	W	W
0106	1885	F	Não	Em que ano foi fundada a Bertrand?	W	W
0107	1886	F	Não	Quem era a locomotiva humana?	W	W
0108	1887	F	Não	Em que concelho fica Loivos do Monte?	W	W
0109	1888	F	Sim	Quem abriu os Jogos Olímpicos de 1948?	W	W
0110	1889	F	Não	O que é M31?	W	W
0111	1890	F	Não	Quando nasceu Machado de Assis?	W	W
0112	1891	D	Não	O que é a Igreja Maronita?	R	R
0113	1892	F	Não	Quanto custou a Mars Observer?	W	W
0114	1893	F	Não	O que é o mascarpone?	X	R
0115	1893	F	Não	De que região provêm?	W	W
0116	1893	F	Não	Em que sobremesa se usa?	W	W
0117	1894	F	Não	Quando foi inaugurado o metropolitano de Lisboa?	W	W
0118	1894	F	Sim	Quantas estações tem agora?	W	W
0119	1895	F	Não	Qual o verdadeiro nome de Michael Caine?	W	W
0120	1896	F	Não	A que altitude está Miguel Pereira?	W	W
0121	1897	F	Não	O que é que Jean Valjean roubou?	W	W
0122	1898	F	Sim	Qual era a divisa austríaca antes de 2002?	W	W
0123	1899	D	Não	Quem foi Monteiro Lobato?	W	W
0124	1899	F	Não	Em que instituição estudou?	W	W
0125	1899	F	Não	O que é que ele estudou?	W	W
0126	1899	F	Não	Quando é que ele criou a Emília?	R	R
0127	1900	F	Não	Quantos municípios tem o distrito do Porto?	W	W
0128	1901	F	Sim	Quem ganhou a NBA em 1995?	W	W
0129	1902	D	Não	O que é a constante de Néper?	U	U
0130	1903	D	Não	O que são os números E?	W	W
0131	1904	F	Não	Qual o período de gestação de ocapi?	W	W
0132	1904	F	Não	Qual o seu peso?	W	W
0133	1905	F	Não	Que concelho fica a leste de Oeiras?	W	W
0134	1906	F	Sim	Que cidades realizaram Jogos Olímpicos antes de 1900?	W	W
0135	1907	F	Não	Diga um recipiente da Grande Cruz do Mérito.	W	W
0136	1908	F	Não	Quando começa o outono no hemisfério sul?	W	W
0137	1909	D	Não	Quem é José Eduardo Pinto da Costa?	W	W
0138	1909	F	Não	Onde é que ele nasceu?	W	W
0139	1910	F	Não	Quantos centímetros há num pé?	W	W
0140	1911	F	Sim	Quem era rei de Portugal em 1860?	W	W
0141	1912	F	Não	Qual a área do Parque Nacional da Peneda-Gerês?	W	W
0142	1913	L	Não	Em que anos foi Nelson Piquet campeão do mundo?	W	W
0143	1914	D	Não	Quem foi Pirro?	W	R
0144	1915	F	Não	Qual o primeiro nome do presidente Trovoada?	W	W
0145	1916	F	Não	Quantas regiões tem Marrocos?	W	W

ID	Grupo	Cat.	R.Temporal	Pergunta	I	II
0146	1916	F	Não	Qual é a maior delas?	W	W
0147	1917	F	Não	Onde é estão sepultados Fernando e Isabel?	R	R
0148	1918	F	Não	Qual é a principal indústria da República Dominicana?	W	W
0149	1919	F	Não	Em que período existiu a República de Veneza?	W	W
0150	1920	F	Não	Onde é que desagua o Ural?	W	W
0151	1920	F	Não	Qual é o seu comprimento?	W	W
0152	1921	F	Não	Qual o comprimento total da rodovia Dom Pedro I?	W	W
0153	1922	F	Não	Quem comandou os gregos em Salamina?	W	W
0154	1923	F	Sim	Quem se tornou rei após a morte de D. Afonso Henriques?	W	W
0155	1924	F	Não	Que ordem foi fundada por Santa Clara?	W	W
0156	1925	D	Não	O que é o Saquê?	R	R
0157	1926	F	Não	Quantos senadores tem o Senado brasileiro actualmente?	W	W
0158	1926	F	Não	Quantos anos tem os seus mandatos?	W	W
0159	1926	F	Não	Quantos senadores tem cada estado?	W	W
0160	1926	F	Sim	Quantos senadores havia em 1972?	R	R
0161	1927	F	Não	Quem foi o último rei da Bulgária?	W	W
0162	1928	F	Não	Diga uma catedral de Sófia.	W	W
0163	1929	F	Não	Em que jornal escreve o Miguel Sousa Tavares?	W	W
0164	1930	F	Não	A que partido pertencia Spadolini?	W	W
0165	1931	L	Não	Quais são as línguas oficiais da Suíça?	W	W
0166	1931	F	Não	E quantos cantões tem?	W	W
0167	1931	F	Não	Quando é que Berna aderiu à Confederação?	W	W
0168	1932	L	Sim	Que selecções disputaram a final da Taça América de 1995?	W	W
0169	1933	F	Não	Quantos espartanos lutaram na Batalha das Termópilas?	W	W
0170	1934	F	Não	Quem sucedeu a Augusto?	W	W
0171	1935	F	Não	Qual é a capital de Timor Ocidental?	W	W
0172	1936	F	Não	Qual o estilo arquitectónico da Torre dos Clérigos?	W	W
0173	1936	F	Não	Quantos degraus tem a torre?	W	W
0174	1936	F	Não	Qual á a altura dela?	W	W
0175	1936	F	Não	Quando foi erigida?	W	W
0176	1937	D	Não	O que era uma trirreme?	R	R
0177	1938	F	Não	Como se chamava o avião supersónico russo?	W	W
0178	1939	F	Não	Qual o túnel ferroviário mais comprido do mundo?	W	W
0179	1939	F	Não	E qual o seu comprimento?	W	W
0180	1939	F	Não	Em que país fica?	W	W
0181	1940	D	Não	O que é o Tux?	R	R
0182	1940	F	Não	Que animal é?	W	W
0183	1940	F	Não	Quem o criou?	W	W
0184	1940	F	Não	Quando?	W	W
0185	1941	F	Não	Diga uma emissora de televisão brasileira que tenha iniciado as suas actividades antes de 1970.	W	W
0186	1942	D	Não	O que é a TVI?	R	R
0187	1943	D	Não	O que é a Union Jack?	W	W
0188	1944	F	Não	Onde é que nasceu o Vítor Baía?	W	W
0189	1944	F	Não	Onde é que começou a jogar futebol?	W	W
0190	1944	F	Não	Quantos golos é que marcou na sua carreira?	W	W
0191	1944	F	Não	Que idade tinha ele quando se mudou para o FC Porto?	W	W
0192	1945	D	Não	Quem é George Vassiliou?	X	X
0193	1946	F	Não	Quero o nome de um vinho húngaro.	W	W
0194	1947	F	Sim	Que clube foi campeão português de voleibol em 1995?	W	W
0195	1947	F	Sim	E quem venceu a Taça de Portugal nesse ano?	W	W
0196	1948	L	Não	Quais são os signos do Zodíaco?	W	W
0197	1949	F	Não	Quantas ilhas tem Cabo Verde?	W	W
0198	1949	F	Não	Quantas são habitadas?	W	W
0199	1949	F	Não	Em qual delas nasceu Eugénio Tavares?	W	W
0200	1949	F	Não	E em qual deles fica a Praia?	W	W

B Segunda Avaliação

Neste apêndice são apresentadas as 200 perguntas que constituíram o conjunto de teste do sistema na segunda avaliação. Para cada pergunta são apresentados o seu identificador, a sua categoria, se tem restrições temporais e o resultado da sua avaliação:

ID	Cat.	Pergunta	
0001	F	Em que cidade se encontra a prisão de San Vittore?	I
0002	F	Onde era o campo de concentração de Auschwitz?	C
0003	F	Quem foi o autor de "Mein Kampf"?	I
0004	F	Qual é a capital da Rússia?	C
0005	F	Quem foi o primeiro presidente dos Estados Unidos?	I
0006	F	Como morreu Jimi Hendrix?	I
0007	F	Com quem se casou Michael Jackson?	I
0008	F	Em que género musical se distingue Michael Jackson?	I
0009	D	O que é a Mossad?	C
0010	F	Quantos crimes são atribuídos ao Monstro de Florença?	I
0011	F	Quantos desempregados há na Europa?	I
0012	F	Quantas religiões monoteístas há no mundo?	I
0013	F	Quantos judeus existem no mundo?	C
0014	F	Quantos detidos há no Corredor da Morte na Califórnia?	C
0015	D	O que é a UNICEF?	C
0016	F	Nomeie uma pessoa acusada de pedofilia.	I
0017	D	Quem é Jean-Bertrand Aristide?	X
0018	F	Mencione um cetáceo.	C
0019	F	Quem escreveu "Ulisses"?	I
0020	F	Onde se situa o CERN?	I
0021	D	Quem é Yves Saint-Laurent?	X
0022	F	Em que dia calha o solstício de verão?	I
0023	F	Onde fica o Museu do Hermitage?	I
0024	F	De que são feitos os cabos de fibra óptica?	I
0025	F	Que forma de governo tem a França?	I
0026	F	Qual o nome da mulher de Kurt Cobain?	X
0027	F	Qual o vulcão activo mais alto da Europa?	I
0028	F	O que significa "Forza Italia"?	I
0029	F	Quando foi lançada a sonda espacial Ulisses?	I
0030	D	O que é a maçonaria?	I
0031	F	Quem foi o último czar da Rússia?	I
0032	F	Qual o acrónimo da Amnistia Internacional?	I
0033	F	Onde se entregam os Óscares?	C
0034	F	Onde fica o arquipélago de Svalbard?	C
0035	F	Onde se realizou a Conferência Mundial da Mulher?	C
0036	F	Indique uma companhia de fast-food.	I
0037	D	Quem foi Rosa Chacel?	I
0038	D	Quem é Christo?	C

ID	Cat.	Pergunta	
0039	D	O que são as FARC?	C
0040	F	Qual a abreviatura do Exército Popular de Libertação do Sudão?	I
0041	F	Em que ano é que o ”War Powers Act”foi aprovado pelo Congresso americano?	I
0042	F	Esmirna fica em que país?	I
0043	F	Qual a localização de Tipaza?	I
0044	F	Quem é o fundador da Motown?	I
0045	F	Como se chama a filha do líder chinês Deng Xiaoping?	I
0046	D	O que é o FSK?	X
0047	F	Em que país fica Vukovar?	I
0048	F	Em que cidade americana se encontra o Museu Warhol?	I
0049	D	Quem é Andy Warhol?	C
0050	F	Quem descobriu o vírus da sida?	I
0051	F	Quem é a ministra do Ambiente alemã?	I
0052	F	Mencione um bonecreiro.	I
0053	D	Quem é o presidente da UEFA?	I
0054	D	O que é a MTV?	I
0055	F	Quem é o líder do KwaZulu?	I
0056	D	O que é a NASA?	C
0057	F	Quem planeou o Palácio dos Desportos São Jorge em Barcelona?	I
0058	F	Em que equipa de básquete joga Shaquille O'Neill?	I
0059	F	Quem é o presidente da Câmara dos Representantes americana?	I
0060	F	Em que ano foi assassinado o presidente chileno Salvador Allende?	I
0061	D	Quem é Marvin Minsky?	I
0062	F	Como se intitula a autobiografia de Nelson Mandela?	I
0063	F	Como se chama a viúva do falecido presidente de Moçambique, Samora Machel?	I
0064	D	Quem é João Havelange?	X
0065	F	O que significa a abreviatura OUA?	I
0066	F	Onde fica Hyde Park?	C
0067	F	Que significa a abreviatura AWACS nos aviões AWACS?	I
0068	F	Onde está preso Hugo Lacour?	I
0069	F	Quantos assinantes tem a MSN?	I
0070	D	O que é a UNICE?	I
0071	F	Quem foi forçado a demitir-se de governador da Caríntia em 1991?	I
0072	F	Qual é a maior empresa industrial da Finlândia?	I
0073	F	Qual o lucro do grupo electrónico e de telecomunicações finlandês Nokia em 1994?	I
0074	D	O que é o CERN?	C
0075	F	Quantos estados-membros tem o CERN?	I
0076	D	Quem é Kevin Mitnick?	I
0077	F	Quando foi criado o CERN?	I
0078	F	Que produz a MCC?	I
0079	F	Qual o monte mais alto do mundo?	I
0080	F	Onde fica Halifax?	C
0081	D	Quem é Umberto Bossi?	I
0082	F	Onde fica o La Scala?	I
0083	F	Onde fica a sede da UNESCO?	C
0084	F	Quem é o realizador de ”Nikita”?	I
0085	F	Quantos anos de residência são necessários para obter a nacionalidade suíça?	C
0086	D	O que é o GIA?	I
0087	F	Qual o cargo de Redha Malek em 1994?	I
0088	F	Quem foi eleito presidente do Conselho Geral da Guiana?	I
0089	F	Que quantia exige o FC Sevilha de Diego Maradona?	I
0090	F	Como se chama o ministro das Finanças polaco?	I
0091	F	Como morreu Juvénal Habyarimana?	I
0092	F	Quem foi derrotado por Andrei Medveded na final do Torneio de Monte-Carlo?	I
0093	F	Qual a nacionalidade do tenista Sergi Bruguera?	I
0094	F	Qual a superfície da República da Chechénia?	I

ID	Cat.	Pergunta	
0095	F	Qual o nome do partido político de Ntsu Mokhele, primeiro-ministro do Lesoto?	I
0096	F	Qual o nome do primeiro-ministro do Ruanda?	I
0097	F	De que país é a escritora Taslima Nasreen?	I
0098	F	Qual o cargo de Albert Reynolds na Irlanda?	I
0099	F	Qual a taxa de desemprego nos Estados Unidos no final de 1994?	I
0100	F	Quando tiveram lugar as eleições europeias de 1994?	I
0101	F	Onde fica a Esfinge de Gizé?	X
0102	F	Onde é Izhevsk?	X
0103	F	Onde fica o Estádio José Alvalade?	C
0104	F	Onde vive José Saramago?	I
0105	F	Onde nasceu Nelson Mandela?	I
0106	F	Qual o maior satélite de Júpiter?	I
0107	F	Onde fica Turku?	C
0108	F	Qual a antiga capital da Polónia?	I
0109	F	Em que distrito fica Paredes de Coura?	I
0110	F	Onde fica Sosnovy Bor?	I
0111	F	Em que ilha fica Ponta Delgada?	I
0112	F	Onde é que nasceu Álvaro Cunhal?	I
0113	F	Onde é o hospital Júlio de Matos?	I
0114	F	Qual é o país mais pequeno da União Europeia?	I
0115	F	Onde fica Gabrovo?	X
0116	F	Qual o estado mais setentrional dos EUA?	I
0117	F	Qual é a capital da Bielorússia?	I
0118	F	Onde desagua o rio Cubango?	I
0119	F	Em que cidade o Mosela encontra o Reno?	I
0120	F	Em que estado do Brasil fica Campo Grande?	I
0121	F	Onde se situa Tianjin?	C
0122	F	Onde é a Ilha do Diabo?	X
0123	F	Quem inventou o saxofone?	I
0124	F	Quem escreveu ”O Principezinho”?	C
0125	F	Quem é o recordista mundial do salto à vara?	I
0126	F	Quem é a ”diva dos pés descalços”?	I
0127	F	Quem é o secretário-geral do PCP?	I
0128	F	De quem é filha Martine Aubry?	I
0129	F	Quem é o Presidente da Câmara de Lisboa?	I
0130	F	Quem é o Presidente da Câmara de Lamego?	I
0131	F	Quem é o embaixador de Portugal em França?	I
0132	F	Com quem casou Whoppi Goldberg?	I
0133	F	Quem foi o primeiro presidente dos Estados Unidos?	I
0134	F	Quem é o ministro-presidente da Renânia-Palatinado?	I
0135	F	Quem foi o último governador de Timor Leste?	I
0136	F	Quem era o marido de Vieira da Silva?	I
0137	F	Quem é o capitão do FC Porto?	I
0138	F	Quem é o imã da mesquita de Lisboa?	I
0139	F	Quem realizou o filme ”Lisbon Story”?	I
0140	F	Quem é a ministra sueca do ambiente?	I
0141	F	Como se chama a rainha da Dinamarca?	I
0142	F	Quem é o padroeiro de Penafiel?	I
0143	F	Que grupo matou Aldo Moro?	I
0144	F	De que grupo é vocalista Teresa Salgueiro?	C
0145	F	Que equipa venceu a Taça CERS em hóquei em patins?	I
0146	F	De que clube é treinador Bobby Robson?	C
0147	F	Que empresa tem uma refinaria em Leça da Palmeira?	I
0148	F	A que partido pertence Duarte Lima?	I
0149	F	Quem financia as IPSS?	I
0150	F	Quantos submarinos tem a marinha portuguesa?	I

ID	Cat.	Pergunta	
0151	F	Quantos municípios há em Portugal?	I
0152	F	Qual o comprimento da Ponte do Freixo?	I
0153	F	Quantos anos tem Inês de Medeiros?	I
0154	F	Qual a distância de Braga a Guimarães?	I
0155	F	Qual a altura do K2?	I
0156	F	Qual o valor da dívida da Eurotunnel?	I
0157	F	Qual a área da Baixa-Saxónia?	I
0158	F	Quantos habitantes tem a República Dominicana?	I
0159	F	Quantos golos marcou Eusébio na sua carreira?	I
0160	F	A que velocidade viaja a luz?	I
0161	F	Quando foram criadas as FPLM (Forças Populares de Libertação de Moçambique)?	I
0162	F	Quando foi a independência de Cabo Verde?	I
0163	F	Quando estreia o filme ”Lisbon Story”?	I
0164	F	Quando foi aprovada a Declaração Universal dos Direitos do Homem?	I
0165	F	Quando morreu Salvador Allende?	I
0166	F	Quando morreu Simão Bolívar?	I
0167	F	Em que dia se comemora a independência do Brasil?	I
0168	F	Quando se tornou ”A Portuguesa”hino nacional?	I
0169	F	Em que ano ocorreu o 25 de Abril?	I
0170	F	Em que embateu o Titanic?	I
0171	F	Qual o símbolo de liderança da Volta a Itália?	I
0172	F	O que foi erguido em 13 de Agosto de 1961?	I
0173	F	A que era alérgico Mel Blanc?	I
0174	F	Que país é campeão do mundo de futebol?	I
0175	F	Como morreu Pasolini?	I
0176	F	Como se tornou o Brasil tetracampeão mundial de futebol?	I
0177	F	Qual foi o primeiro filme sonoro português?	I
0178	F	Que vende Fausto ao Diabo?	I
0179	F	Que animal é o símbolo da Namíbia?	I
0180	F	Qual o pseudónimo de Álvaro Cunhal?	I
0181	F	Qual é a nacionalidade de Yordan Letchkov?	I
0182	F	Qual a nacionalidade de Hercule Poirot?	I
0183	F	O que era Napoleão III a Napoleão Bonaparte?	I
0184	F	De que material são os frisos do Parténon?	I
0185	F	Qual a patente de Alfred Dreyfus?	I
0186	F	De que cor é a neve?	I
0187	F	Qual é a moeda iraquiana?	I
0188	F	Qual o endereço da Livraria Barata?	I
0189	D	Quem é Leonor Beleza?	I
0190	D	Quem é Arnold Ruutel?	I
0191	D	Quem é Wim Duisenberg?	X
0192	D	Quem é Rocha Vieira?	C
0193	D	Quem é Guilherme da Fonseca?	I
0194	D	Quem é Fernando Gomes?	X
0195	D	Quem é Valentina Terechkova?	I
0196	D	Quem é Jorge Amado?	C
0197	D	O que é o PC do B?	I
0198	D	O que é o PSN?	I
0199	D	O que é o CSKA?	C
0200	D	O que é a Vigor?	I