UNIVERSIDADE DE LISBOA

Faculdade de Ciências

Departamento de Informática

INFRA-ESTRUTURA DE UM SERVIÇO ONLINE DE RESPOSTA-A-PERGUNTAS COM BASE NA WEB PORTUGUESA

Lino Miguel Silva Rodrigues

Mestrado em Engenharia Informática

2007

UNIVERSIDADE DE LISBOA

Faculdade de Ciências

Departamento de Informática

INFRA-ESTRUTURA DE UM SERVIÇO ONLINE DE RESPOSTA-A-PERGUNTAS COM BASE NA WEB PORTUGUESA

Lino Miguel Silva Rodrigues

Projecto orientado pelo Prof. Dr. António Horta Branco

Mestrado em Engenharia Informática

2007

Declaração

Lino Miguel Silva Rodrigues, aluno nº 28143 da Faculdade de Ciências da Universidade de Lisboa, declara ceder os seus direitos de cópia sobre o seu Relatório de Projecto em Engenharia Informática, intitulado "Infra-estrutura de um Serviço Online de Resposta-a-Perguntas com base na Web Portuguesa", realizado no ano lectivo de 2006/2007 à Faculdade de Ciências da Universidade de Lisboa para o efeito de arquivo e consulta nas suas bibliotecas e publicação do mesmo em formato electrónico na Internet.

FCUL, 20 de Julho de 2007

António Horta Branco, supervisor do projecto de Lino Miguel Silva Rodrigues, aluno da Faculdade de Ciências da Universidade de Lisboa, declara concordar com a divulgação do Relatório do Projecto em Engenharia Informática, intitulado "Infra-estrutura de um Serviço Online de Resposta-a-Perguntas com base na Web Portuguesa".

FCUL, 20 de Julho de 2007

Resumo

A Internet promoveu uma nova forma de comunicação global, com um impacto profundo na disseminação da informação. Em consequência, tornam-se necessárias novas soluções tecnológicas que permitam explorar os recursos actualmente disponíveis. Numa era em que os motores de busca de documentos já fazem parte da vida quotidiana do cibernauta, o próximo passo é permitir que os utilizadores da rede obtenham breves respostas a perguntas específicas.

O projecto QueXting foi encetado pelo Grupo de Fala e Linguagem Natural (NLX) do Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa com o objectivo principal de contribuir para um melhor acesso à informação, possibilitando a realização de perguntas em Português e a obtenção de respostas a partir da informação disponível nos documentos escritos em língua portuguesa. Para tal, pretende oferecer livre acesso a um serviço *online* que processa os documentos da Internet escritos nesta língua e que começará por proporcionar respostas a perguntas factuais.

O sistema de resposta-a-perguntas QueXting tem como pilares a arquitectura e metodologia recentemente amadurecidas neste domínio científico e diversas ferramentas linguísticas, específicas para a língua portuguesa, que o NLX tem vindo a desenvolver. O processamento linguístico específico é um dos factores chave que distingue a tarefa de resposta-a-perguntas das restantes tarefas de recuperação e extracção de informação, permitindo um processamento profundo dos pedidos de informação e a obtenção de respostas exactas.

Esta dissertação apresenta os resultados do desenvolvimento da infra-estrutura do sistema QueXting, que servirá de base ao processamento específico de diversos tipos de perguntas factuais. Apresenta ainda os resultados obtidos no processamento de um tipo de pergunta específico, para o qual foram realizadas avaliações preliminares.

PALAVRAS-CHAVE:

Resposta-a-Perguntas, Recuperação e Extracção de Informação, Processamento de Linguagem Natural, Processamento Computacional do Português.

Abstract

The Internet promoted a new form of global communication, with deep impact in the dissemination of information. As a consequence, new technological solutions are needed for the exploitation of the resources thus made available. At a time when document search engines are already part of the daily life of Internet users, the next step is to allow these users to obtain brief answers to specific questions.

The QueXting project was undertaken by the Natural Language and Speech Group (NLX) at the Department of Informatics of the Faculty of Sciences of the University of Lisbon, with the main goal of contributing to a better access to information available in the Portuguese language. To this end, a web service supporting questions in Portuguese should be made freely available, gathering answers from documents written in this language.

The QueXting question-answering system is implemented through a general methodology and architecture that have recently matured in this scientific domain. Furthermore, it is supported by several linguistic tools, specific for Portuguese, that the NLX group has been developing. This specific linguistic processing is a key factor distinguishing the task of question-answering from the remaining tasks of information retrieval and extraction, allowing the deep processing of information requests and the extraction of exact answers.

This dissertation reports on the development of the infra-structure of the QueXting system, which will support the specific processing of several types of factoid questions. Such processing has already been applied to one specific type of question, for which some preliminary evaluations were completed.

KEYWORDS:

Question answering, Information Retrieval, Information Extraction, Natural Language Processing, Portuguese, Computational Processing of Portuguese

Agradecimentos

Neste ano de adaptação a novas realidades pessoais e profissionais, certas pessoas tiveram uma presença determinante. Aproveito esta oportunidade para deixar os meus sinceros agradecimentos a todos eles.

Ao Professor António Branco, pela orientação e pela união que fomenta no grupo NLX. Aos membros do NLX: João Silva, Pedro Martins, Eduardo Ferreira, Filipe Nunes, Mariana Avelãs, Rosa Del Gaudio, Francisco Costa e Marcos Garcia. À FCT pelo financiamento do projecto e ao DI-FCUL pela disponibilização do espaço. À Linguateca, pela disponibilização dos recursos de avaliação da conferência CLEF.

A toda a minha família, pelo espaço concedido para que eu pudesse terminar este trabalho. Ao Pantufa, por lembrar-me que *casa* é onde estão as nossas tarefas e laços mais importantes. Ao Leo, Hugo Marrão, Ana Cruz, Raquel e Cátia – pelo apoio e boas memórias. Ao Dino e Amílcar, pela companhia na FCUL. Ao Tiago e a Hideaki Anno por uma obra única, e aos *Tool* por *Lateralus*.

A todas as crianças, pelas perguntas mais difíceis

Índice

Índice de Figuras e Tabelas

Capítulo 1

Introdução

1.1 Enquadramento institucional

Esta dissertação foi realizada no ano lectivo de 2006/2007, no âmbito do Mestrado em Engenharia Informática da Faculdade de Ciências da Universidade de Lisboa (FCUL). O trabalho apresentado neste documento foi levado a cabo no Grupo de Fala e Linguagem Natural (Grupo NLX) do Departamento de Informática da FCUL, enquadrando-se no projecto de I&D *QueXting*.

O NLX realiza actividades de investigação e desenvolvimento nos domínios de Inteligência Artificial e Ciência Cognitiva, focando-se particularmente no Processamento de Linguagem Natural. O grupo tem vindo a desenvolver diversos recursos e ferramentas linguísticas para o processamento computacional da língua portuguesa. O QueXting é um projecto financiado pela Fundação para a Ciência e Tecnologia, sob o contrato POSI/PLP/61490/2004.

1.2 Enquadramento tecnológico e científico

Informação na rede

O crescimento da Internet ofereceu novas oportunidades de acesso a informação e serviços, mas fez surgir novas necessidades para a exploração dessas oportunidades. Constituindo a informação espalhada pelo ciberespaço um enorme repositório de conhecimento, estas necessidades incluem ferramentas para lidar com grandes quantidades de dados de forma relevante e atempada. Os actuais motores de busca (e.g. Yahoo, Google[1]) constituem-se como os instrumentos actualmente mais utilizados para responder a estas necessidades, permitindo encontrar documentos que poderão ser relevantes dadas as palavras-chave introduzidas pelo utilizador.

1 Os endereços de Internet destes motores e doutros mencionados adiante estão disponíveis no Anexo A.

Estas aplicações não foram as primeiras ferramentas implementadas para facilitar o acesso à informação. Na realidade, as primeiras ferramentas com esta finalidade foram os directórios, que apenas agrupam endereços de páginas por tópicos e deixam o resto da busca nas mãos do utilizador. Os motores de busca de documentos vieram facilitar esta tarefa ao permitir encontrar documentos relevantes dado o conjunto de palavras-chave introduzidas.

As Figuras 1 e 2 exemplificam a interface, provavelmente familiar para o leitor, de uma destas aplicações. É possível verificar como, para pedidos genéricos de informação sobre um determinado tema, são obtidas referências interessantes, e até um pequeno excerto do conteúdo desses documentos.

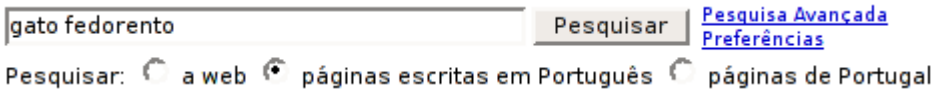

Figura 1: Caixa de pesquisa num motor de recuperação de documentos

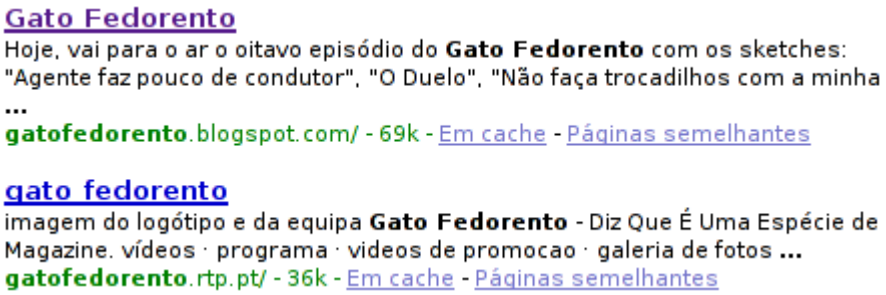

Figura 2: Excerto da lista de resultados relevantes

Para encontrar as respostas de muitas perguntas, um utilizador terá que realizar uma tarefa complexa pelas suas próprias mãos: a de seleccionar e ler vários documentos de entre os que parecerem mais promissores. Infelizmente, mesmo os documentos relevantes (cujo assunto se refere realmente à pesquisa feita) podem não conter uma resposta. Por exemplo, o primeiro documento devolvido pelo Google à pesquisa *"Quem são os Gato Fedorento?"* é o blogue desse grupo de comediantes, mas em nenhuma parte desse blogue é dada uma resposta específica.

No entanto, é frequente o cibernauta precisar de respostas para questões específicas, como

"Quem foi a segunda pessoa a pisar a Lua?" ou *"Qual é a diferença entre ovos de galinha brancos e castanhos?"*. Questionar é uma forma natural de expressão e raciocínio do ser humano que não se traduz simplesmente em palavras-chave e que não fica respondida com uma lista de documentos.

Avanços em Resposta-a-Perguntas

Nos últimos anos, vários projectos de I&D têm formulado técnicas para Sistemas de Resposta a Perguntas (SRP) que obtêm respostas concretas, em vez de listas de documentos, como resultado de uma pergunta formulada em linguagem natural. Protótipos que participam em competições como a TREC/QA (*Text Retrieval Conference / Question Answering track*) têm vindo a mostrar progressos neste domínio. Para as questões mais simples, de carácter factual, alguns destes sistemas já atingem bons resultados. As arquitecturas e métodos utilizados nestes sistemas podem considerar-se estáveis e agora muitos investigadores e as próprias conferências de avaliação começam a atacar seriamente tipos de questões mais complicados.

Apesar de este ser um campo que está há muito tempo em desenvolvimento, desde as primeiras aplicações da Inteligência Artificial, a resposta-a-perguntas automática em domínio aberto (sem restrição de tema) é uma tecnologia que só na última década tem dado frutos. Actualmente, alguns protótipos *online* já foram adquiridos por empresas (Ask, Brainboost); as empresas que detêm os motores de busca de documentos começam a dar alguma atenção ao assunto [36]; e muito recentemente formou-se uma nova empresa (Powerset) dedicada à resposta-a-perguntas na Internet, com o objectivo explícito de destronar o Google como porta de acesso privilegiada à rede global.

Havendo SRP de acesso livre e sem restrição de domínio cuja base de conhecimento é a *Web* (cf. lista no Anexo A), um aspecto relevante a focar é o de que essas aplicações são centradas na língua inglesa. Mesmo quando é permitido ao utilizador introduzir perguntas em Português (por exemplo, no caso do sistema AnswerBus), as questões são traduzidas automaticamente para Inglês e é a rede de documentos em Inglês que é pesquisada para se obter respostas.

A questão da língua é um problema inexistente na busca de documentos: um motor de busca é em grande medida independente dos idiomas visto que os algoritmos que usa são apropri-

ados para encontrar documentos escritos em qualquer língua. Porém, um SRP usa ferramentas de Processamento de Linguagem Natural (PLN) que o restringem a encontrar respostas a partir de documentos escritos em idiomas com os quais esses componentes PLN possam lidar.

Tal como a maioria das redes de idiomas diferentes do Inglês, a rede portuguesa ainda não se encontra acessível através de RP tal como é já tecnologicamente viável. O projecto QueXting do grupo NLX pretende colmatar essa lacuna, proporcionando um melhor acesso à informação pelos utilizadores da rede portuguesa. Para tal, pretende oferecer livre acesso a um serviço *online* que proporcione respostas a perguntas factuais formuladas em Português, com base na informação contida na rede dos documentos escritos nesta língua. Este sistema será a primeira aplicação deste tipo a focar-se directamente nos documentos da Internet escritos na língua portuguesa.

1.3 Objectivos

O trabalho realizado no âmbito desta dissertação visa a implementação da infra-estrutura projectada para o sistema QueXting. A partir desta primeira versão de base, novas ferramentas e capacidades foram e serão adicionadas ao sistema. Entretanto, diferentes tipos de perguntas serão alvo de processamento específico. O conhecimento linguístico envolvido é suportado por ferramentas e recursos linguísticos desenvolvidos no grupo NLX, enquanto que a recolha de documentos-fonte se apoia nos motores de busca existentes *online*.

Com esta aplicação pretende-se, em traços gerais, tirar partido de toda a informação que se encontra disponível na rede e indexada por motores de busca, proporcionar uma interface em língua portuguesa para toda essa informação e contribuir para o avanço da área de RP no panorama científico português.

O objectivo principal de concretização da infra-estrutura básica do QueXting implicou um estudo da arquitectura prevista e a implementação dos módulos que a constituem. Os diferentes módulos resolvem partes complementares do problema, desde o processamento da questão até à selecção de respostas, passando pela recolha de documentos-fonte.

Alguns destes módulos socorrem-se de ferramentas linguísticas, algumas delas específicas para o Português. Consequentemente, um sub-objectivo consiste na adaptação das ferramentas e recursos linguísticos que já existem (e.g. segmentador de frases, ontologias), e no desenvol-

vimento de outros que ainda não estão concretizados (e.g. dicionário de perguntas/respostas).

Outro objectivo importante consiste na disponibilização do serviço, ou seja, na colocação da aplicação *online* através de uma página *web.* Este objectivo implica uma avaliação e optimização da qualidade do serviço e dos resultados obtidos nas pesquisas. À medida que se iniciou o desenvolvimento tornou-se premente a definição de mecanismos adequados de avaliação do sistema, que permitirão guiar o seu desenvolvimento.

1.4 Estrutura da dissertação

Após esta introdução ao tema e aos objectivos do projecto, os próximos capítulos irão abordar em pormenor o contexto do trabalho (Capítulo 2), a arquitectura do QueXting (Capítulo 3) e a avaliação (Capítulo 4).

O segundo capítulo – *Contexto* – dará uma perspectiva científica e histórica da área; abordará as características dos SRP segundo factores como tipos de utilizadores/pedidos, domínio de aplicação, fontes de respostas, recursos; discriminará os tipos de SRP segundo as capacidades possibilitadas pelas características descritas acima; e fará uma análise aos SRP disponíveis na rede e aos SRP em língua portuguesa.

O terceiro capítulo – *O Sistema QueXting* – irá descrever a abordagem de desenvolvimento, o processo de resolução de pedidos e os recursos utilizados pela aplicação. Em seguida, irá descrever pormenorizadamente cada um dos módulos do sistema.

O quarto capítulo – *Avaliação* – abordará as conclusões alcançadas sobre as necessidades essenciais para avaliação: recursos, critérios e métricas. Finalmente, são descritas as medidas tomadas para avaliação do QueXting e os resultados das avaliações preliminares efectuadas.

O capítulo final – *Conclusão* – irá resumir os principais produtos e resultados da pesquisa e do trabalho levados a cabo. Serão ainda abordados os aspectos que poderão ser melhorados ou desenvolvidos no futuro. A dissertação fica concluída com um breve comentário crítico envolvendo aspectos transversais a todo o trabalho.

Capítulo 2

Contexto

Este capítulo apresenta uma panorâmica do contexto em que se insere o trabalho realizado, dos pontos de vista científico e tecnológico. São descritas as características gerais dos utilizadores e seus pedidos, assim como as características dos próprios sistemas ao nível de capacidades, recursos e processos. Finalmente, são abordadas aplicações específicas relevantes para o sistema desenvolvido no âmbito desta tese: os SRP baseados na Internet e os SRP em língua portuguesa.

2.1 Fundações dos Sistemas de Resposta-a-Perguntas

O cruzamento entre o domínio da informação linguística e o domínio do processamento computacional de informação pode ser encarado de duas perspectivas complementares. De uma perspectiva, está em evidência a vantagem de interagir com os computadores de uma forma mais natural: através da linguagem humana. De outra perspectiva, jaz a possibilidade (ou necessidade) de processar quantidades potencialmente vastas de dados (texto, áudio) linguísticos salvaguardados em formato digital.

As aplicações que se posicionam sob a primeira perspectiva são do domínio da **Interacção** em Linguagem Natural (ILN), enquanto as ferramentas desenvolvidas sob a segunda perspectiva pertencem ao domínio do **Processamento** de Linguagem Natural (PLN). Muitas das ferramentas de PLN servem de base ou apoio a aplicações concretas centradas no utilizador – as aplicações de ILN. Este tipo de reutilização (ou construção de funcionalidades práticas de alto nível com base em capacidades anteriormente conseguidas) é um dos pilares do trabalho desenvolvido na área de resposta-a-perguntas.

Por configuração própria da área, as ferramentas de PLN aplicam-se à linguagem natural escrita ou a uma representação discreta da fala. A linguagem escrita é representada interna-

mente pelos computadores como sequências de códigos, em que cada código representa um caracter, isto é, um dos símbolos básicos: letra, dígito, ou símbolo de pontuação. Com o desenvolvimento dos métodos de PLN, os textos disponíveis em meios digitais podem ser processados de uma forma prática e útil para aplicações de ILN. Uma das aplicações que beneficia dessas capacidades e simultaneamente motiva o seu desenvolvimento é a dos SRP, que têm vindo a ganhar relevo nos últimos anos.

2.1.1 Perspectiva histórica

A origem dos SRP remonta aos primeiros desenvolvimentos da Inteligência Artificial. Os primeiros sistemas deste tipo foram desenvolvidos na década de 1960 e serviam de interface para sistemas periciais dedicados a domínios específicos [36]. Um artigo de Simmons (1965) revê 15 sistemas construídos nos 5 anos precedentes – aplicações conversacionais, interfaces (*front-ends*) para repositórios de dados e sistemas que procuram respostas em fontes de texto, como enciclopédias [27, 12]. Desta primeira vaga de SRP, destacam-se os seguintes:

- BASEBALL: um programa para responder a questões sobre jogos de basebol jogados na liga americana durante um ano. [10, 12]
- STUDENT e SHRDLU: criados para resolver problemas, respectivamente nos domínios de álgebra e mundo de blocos. [37, 11]
- LUNAR: demonstrado em 1971 na *Second Annual Lunar Conference*, respondeu correctamente a 78% das perguntas acerca das rochas lunares. [38, 30]

Nas décadas de 1970/80, foram desenvolvidas teorias mais abrangentes em linguística computacional que levaram a projectos ambiciosos de compreensão de texto e SRP. Uma teoria sobre RP foi proposta por Lehnert e implementada no sistema QUALM, com uma taxonomia de questões [17]. As questões foram classificadas em categorias conceptuais que eram usadas para análise inferencial através da teoria de dependência conceptual de Schank. Segundo Schank, a base da linguagem é conceptual, independente do idioma, e uma estrutura (*framework*) de representação de significado facilitaria o raciocínio sobre o conteúdo de um discurso [26, 11].

Numa outra orientação do desenvolvimento em SRP, foi colocado de lado o conhecimento específico e a capacidade dos sistemas compreenderem texto [30]. Estes sistemas não depen-

dem de uma base de conhecimento especificamente criada para o efeito e, portanto, são independentes do domínio. Em lugar de estarem ligados aos paradigmas e métodos de Inteligência Artificial, usam técnicas de recuperação de informação com o objectivo de obter passagens-resposta a partir de um corpo de texto de grandes dimensões. O sistema MURAX de Kupiec (1993) usa uma enciclopédia *online* como fonte de respostas para as perguntas que Kupiec definiu como sendo de classe fechada (perguntas de resposta curta, baseadas em factos) [16, 11].

Os sistemas referidos partilham a característica de se basearem em repositórios de dados estáticos. No entanto, é possível utilizar fontes dinâmicas, das quais a Internet é um exemplo paradigmático. O sistema desenvolvido no âmbito desta tese é um sistema deste tipo e, como tal, uma abordagem mais pormenorizada aos SRP baseados na Internet será feita mais adiante neste capítulo.

Actualmente estão em desenvolvimento vários sistemas independentes de domínio que usam grandes colecções de textos (quer estáticas, quer dinâmicas) e combinam várias técnicas de PLN para encontrar respostas a diversos tipos de perguntas. A utilização de grandes colecções de texto foi impulsionada não só pelo crescimento da informação disponível em meios digitais e do número de utilizadores com acesso a ela [22], mas também pela inclusão da tarefa de RP em conferências de avaliação como as descritas nos parágrafos seguintes.

2.1.2 Conferências de avaliação

Em 1999 surgiu o primeiro fórum de avaliação de SRP, incluído na *Text Retrieval Conference* (TREC). Os sistemas que participam nesta competição devem responder a perguntas de qualquer domínio, procurando respostas num corpus estático de texto. Este evento propiciou a I&D em SRP de domínio aberto baseados em grandes colecções de texto. Um pequeno estudo do âmbito deste evento nos últimos anos está disponível no Anexo B. O melhor sistema concorrente em 2005 foi capaz de responder correctamente a 71% das perguntas factuais testadas [35].

Devido ao facto de a TREC se dedicar essencialmente ao processamento da língua inglesa, surgiu uma nova conferência de avaliação, o *Cross Language Evaluation Forum* (CLEF). que aborda o campo de RP desde 2003.[2] Além de focar explicitamente linguagens diferentes do In-

2 Outra conferência que se especializa no cruzamento de idiomas é a NTCIR, que se realiza na Ásia.

glês, este evento dedica atenção ao cruzamento entre as línguas, abordando por exemplo a possibilidade de realização de perguntas numa língua e obtenção de respostas noutra. Para o Português, tem sido utilizado um corpus contendo artigos dos jornais Público e Folha de São Paulo de 1994 e 1995. O melhor sistema nesta língua em 2006 foi o da Priberam, com 65% de acerto nas perguntas factuais testadas [20].

2.2 Características dos SRP

Acima de tudo, um SRP é uma interface para um serviço que proporciona um retorno adequado ao pedido efectuado: permite ao utilizador fornecer ou introduzir uma pergunta específica, com o objectivo de obter uma resposta adequada. No entanto, o núcleo do sistema envolve vários componentes que efectuam o processamento do pedido e das fontes de respostas, empregando diversos métodos de processamento sintáctico, semântico ou pragmático. O papel do sistema é interpretar a pergunta, procurar a informação pedida numa colecção de textos, base de conhecimento ou repositório de dados e identificar os excertos, frases ou expressões que efectivamente consistem em respostas.

Portanto, um sistema deste tipo lida com duas entidades fundamentais: o utilizador (cujo pedido deve ser interpretado da melhor forma possível, possibilitando um retorno adequado), e a fonte em que se baseia para obter as respostas que o utilizador procura. Ao longo do processo de obtenção de respostas, existe ainda a necessidade de aplicar métodos ou usar recursos específicos de PLN.

Os restantes parágrafos desta Secção abordam individualmente estas facetas dos SRP, ficando para a Secção 2.3 uma classificação dos SRP com base nas capacidades e recursos utilizados.

2.2.1 Utilizadores e pedidos

A interface pessoa-máquina de um sistema deste tipo é variável, dada a diversidade de utilizadores que poderão interagir com ele. Utilizadores mais avançados poderão realizar pedidos de informação mais complexos, pelo que a possibilidade de introduzir contexto ou pistas para o sistema não deve ser menosprezada. Daí as conferências de avaliação destes sistemas começarem a introduzir aspectos de interactividade. Um exemplo de interactividade é a clarificação

de contexto pelo utilizador, por iniciativa própria ou após um pedido de clarificação efectuado pelo sistema. Outro exemplo é o dos sistemas que respondem a perguntas no contexto das interacções anteriores com o utilizador (e.g. sequência de questões, com progressiva especificidade, sobre um determinado evento).

No entanto, nem todos os pedidos de informação são tão complexos. Eis alguns dos casos de uso típicos do sistema por parte dos seus utilizadores:

- perguntas casuais por utilizadores comuns;
- procura de preços e características de produtos por potenciais compradores;
- pesquisas relacionadas com estudos de mercado/negócio por analistas empresariais;
- procura de dados pormenorizados por investigadores profissionais.

Consoante as técnicas e os recursos utilizados, um SRP terá capacidade para responder a diferentes tipos de perguntas [34, 11]:

Tipo de pergunta	***Exemplos***
Factual (de resposta directa)	*Quem foi a primeira pessoa a pisar a Lua?*
Lista (exigindo uma compilação de vários itens)	*Que modelos de automóveis produz a BMW?*
Definição (de resposta mais aberta e, em consequência, potencialmente mais extensa);	*Quem é Durão Barroso?* *O que é âmbar?*
Especulativa (cuja resposta pode não estar presente em nenhum documento e exige algum método de raciocínio)	*A indústria aérea está em apuros?* *Porque existe uma crise política em Lisboa?*

Tabela 1: Tipos de perguntas

Os níveis de sofisticação de utilizadores e perguntas foram resumidos por um conjunto de investigadores ligados às *Document Understanding Conferences,* nas quais se insere a TREC, através do diagrama da Figura 3, originalmente apresentado em [6]. O parágrafo após a Figura 3 descreve os tipos de utilizadores e pedidos representados nesse diagrama.

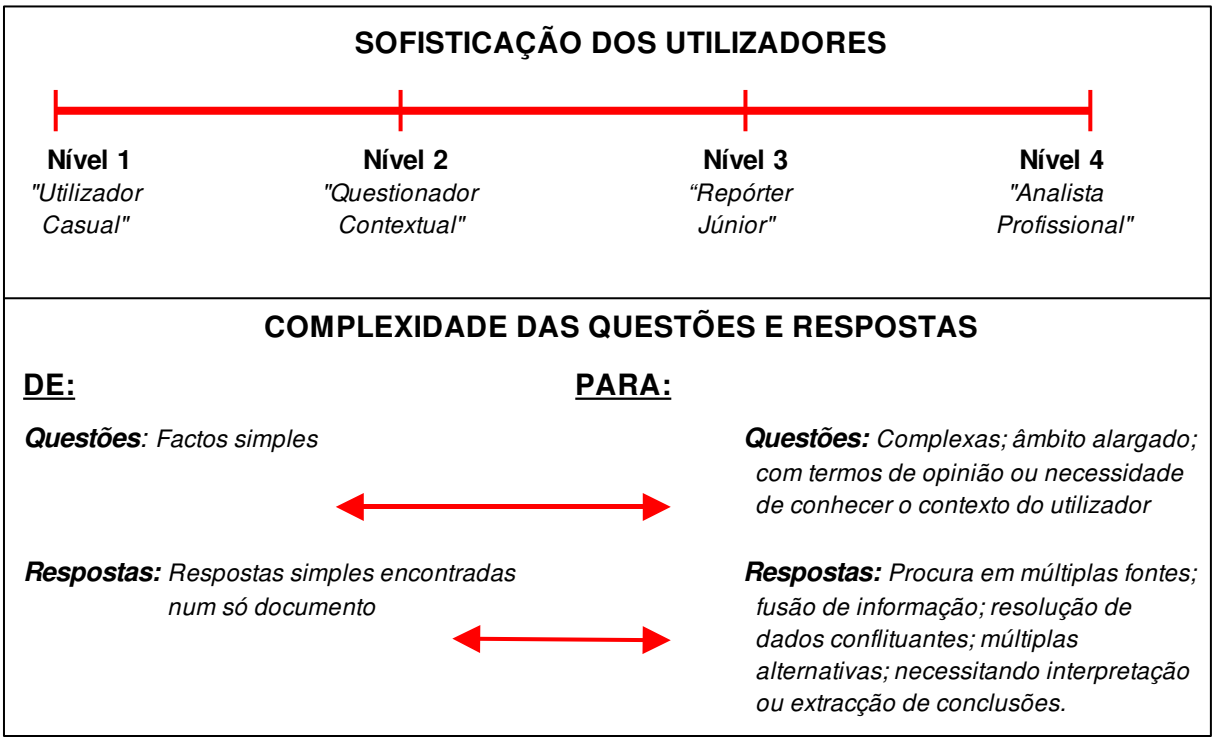

Figura 3: Níveis de sofisticação de utilizadores e perguntas

O *Utilizador Casual* é aquele que realiza perguntas simples de carácter factual, cuja resposta poderá estar numa frase curta. O denominado *Questionador Contextual* é o tipo de utilizador para o qual o sistema pode fornecer certos modelos que o utilizador preenche com informação de contexto relevante para a pergunta. O *Repórter Júnior* representa o utilizador que pretende realizar pesquisas ainda de pouca profundidade sobre um determinado assunto, ou pesquisas sobre um aspecto particular de um tópico mais geral; trata-se de um utilizador focado em factos mas que potencialmente precisa de cruzar informação de diversas fontes. Finalmente, o perfil de *Analista Profissional* (que inclui repórteres seniores, investigadores policiais ou judiciais, analistas financeiros, cientistas) cobre os casos de questionadores com necessidades de informação de carácter mais profundo.

2.2.2 Fontes de respostas e domínio de aplicação

Como referido anteriormente, algumas aplicações iniciais de RP constituíam interfaces para sistemas periciais e, como tal, consultavam uma base de conhecimentos construída por especialistas. Outras interfaces deste tipo podem ser usadas para obter informação a partir de repositórios de dados ou manuais técnicos. Estes tipos de SRP são sistemas de domínio fecha-

do (restrito) que usam uma fonte de respostas também fechada (estática). No entanto, nem sempre estas características surgem associadas, como a Tabela 2 demonstra:

Tipos de sistemas	***Domínio de aplicação***	***Fontes de respostas***
Interfaces para sistemas periciais e bases de dados	Restrito	Estáticas
Baseados em enciclopédias	Restrito ou aberto, dependendo do tema da enciclopédia	Geralmente estáticas (exceptuando enciclopédias *online* editáveis, como a Wikipédia)
Baseados em colecções de texto não estruturadas (embora possivelmente compiladas de forma organizada)	Restrito ou aberto, dependendo da abrangência dos documentos	Estáticas (colecções locais de documentos) ou dinâmicas (e.g. a Internet)

Tabela 2: Fontes de respostas e domínio de aplicação de um SRP

A vantagem dos sistemas de domínio restrito é que, apesar de responderem apenas a perguntas sobre um tema fixo, têm um modelo desse domínio representado em regras ou componentes da base de dados que permitem a utilização de técnicas avançadas como prova de teoremas ou métodos complexos de raciocínio [22].

No entanto, o foco da I&D actual, influenciado pela crescente quantidade de informação disponível em meios digitais e pelas conferências de avaliação que surgiram na última década, centra-se na utilização de grandes colecções de textos (e.g. compilações de publicações jornalísticas, Internet). Grandes colecções tendem a melhorar o desempenho, desde que os temas incluídos nas colecções digam respeito ou incluam o domínio das perguntas. A abrangência destas colecções torna possível o desenvolvimento de sistemas independentes de domínio. Nestes sistemas, que não são construídos em torno de uma base de dados ou regras específicas, é fácil substituir ou complementar as fontes de respostas.

A presença de informação redundante, que pode ocorrer facilmente em colecções de grande escala, é um factor determinante. A possibilidade de dispor da mesma informação expressa de forma diferente, em diferentes contextos e documentos, facilita o sucesso das técnicas de PLN usadas para captar padrões e significado nos textos. Além disso, torna possível realizar um cruzamento das respostas obtidas em diferentes contextos que favoreça as respostas que

ocorrem mais vezes, contribuindo assim para eliminar eventuais falsos positivos, i.e. respostas incorrectas [36].

2.2.3 Processo de resolução

Esta secção descreve brevemente os processos que ocorrem entre a formulação da questão por parte de um utilizador e a obtenção de uma resposta pelo sistema, focando-se na perspectiva dos SRP baseados em colecções de texto (como é o QueXting). O diagrama apresentado na Figura 4 mostra como se enquadram estes processos numa arquitectura básica típica:

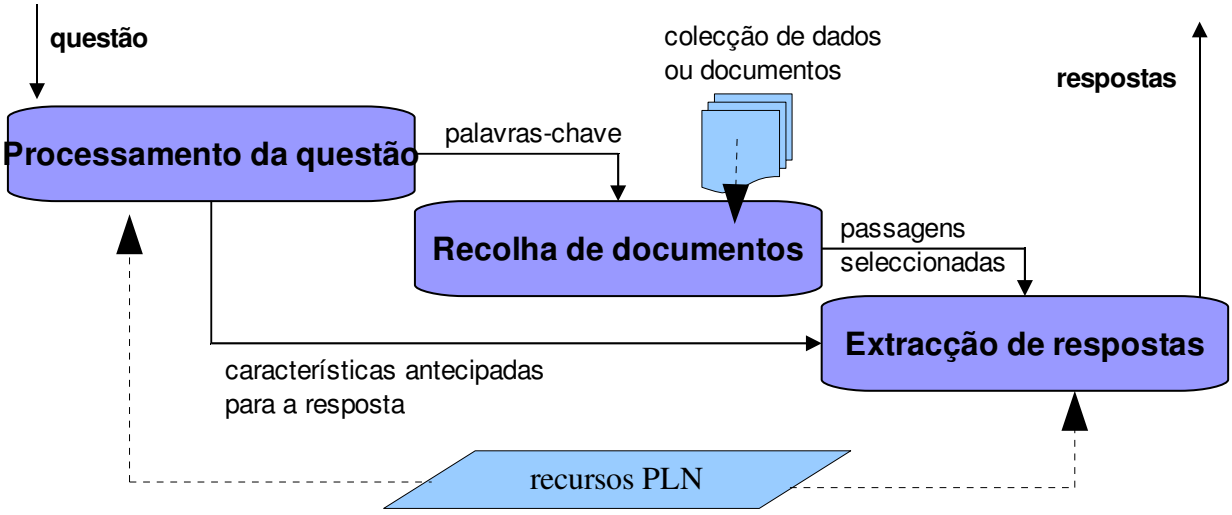

Figura 4: Processos e arquitectura típica

Considerando a pergunta de um utilizador como o problema que o sistema deve resolver, o primeiro passo na resolução do problema consiste na interpretação da pergunta. As perguntas têm uma semântica inerente que os humanos compreendem e os SRP conseguem processar até certo ponto. O aspecto mais importante desta semântica é o tipo de resposta esperado, que pode ser determinado pelo próprio padrão sintáctico da pergunta – em que se inclui o pronome interrogativo usado – e pelo tipo de entidades referidas.

Outro passo essencial na resolução do problema é determinar quais são os documentos (ou entradas) onde poderá estar contida uma resposta adequada. Para isto, são usadas técnicas de recuperação de informação, nomeadamente recorrendo a motores de busca de documentos (e/ou ao pré-processamento/indexação de documentos) e usando palavras-chave encontradas

no passo anterior.

O último passo corresponde à identificação de fragmentos de texto que contenham as características previstas, não só ao nível de palavras-chave como do tipo (semântico) das entidades referidas. Uma eventual capacidade de identificar respostas exactas, em lugar de simples excertos do texto, facilita a tarefa do utilizador, mas pode levar ao aumento da ocorrência de respostas que, desprovidas de qualquer contexto, se revelem insuficientes ou mesmo incorrectas.

Actualmente, os SRP incluem estes três passos como módulos de uma arquitectura que, pelo menos para as perguntas factuais, já consegue produzir resultados significativos. Como referido na Secção 2.1.2, os resultados medidos nas conferências de avaliação podem chegar a 71% de perguntas factuais correctamente respondidas [35].

Dada a variedade e importância dos recursos de PLN utilizados, uma abordagem a este aspecto será realizada na secção seguinte. Em sistemas mais complexos, que tenham características de interactividade ou possibilidades de fornecer contexto, os 3 módulos indicados e os recursos de PLN neles utilizados são complementados com capacidades adicionais, nomeadamente a construção de modelos do utilizador/domínio ou bases de conhecimento *ad hoc*, para resolver os pedidos.

2.2.4 Recursos e métodos de PLN que sustentam os SRP

O processo de RP partilha alguns métodos de PLN com outras tarefas relacionadas: as tarefas de Recuperação de Informação (RI) e de Extracção de Informação (EI). Para identificar documentos onde poderão encontrar respostas, os SRP usam métodos de RI. Para reconhecer entidades nomeadas usam métodos de EI.

Em traços gerais, os sistemas de RI permitem localizar documentos relevantes para um determinado pedido (*query*), deixando ao utilizador a tarefa de percorrer (e eventualmente explorar individualmente) cada resultado da lista ordenada de documentos. Quanto aos sistemas de EI, estes permitem identificar a informação textual relevante, dada uma representação (*template*) do evento ou entidade alvo e operando sobre um conjunto de documentos considerando relevante para a tarefa.

Apesar de estas tarefas serem complementares, a EI depende de regras ou padrões com

bastante conhecimento do domínio a que se aplica, pelo que a simples combinação das técnicas de RI e EI seria inadequada para responder a questões em domínio aberto: seriam necessárias regras de extracção para todos os domínios possíveis e o tipo de perguntas seria restringido pelo tipo de *templates* definidos. Além disso, a identificação de uma resposta precisa é um processo mais exigente e específico do que as tarefas de RI e EI. Portanto, os SRP usam outras técnicas de PLN que lhes permitem capturar pelo menos parte da semântica das questões e efectuar unificações léxico-semânticas das questões com as potenciais respostas. [11]

Os métodos de PLN utilizados podem ser de âmbito linguístico profundo ou superficial. A necessidade de resolver os pedidos de um utilizador em tempo útil é um factor que pode determinar uma combinação híbrida de métodos de PLN. A título de exemplo: se por um lado é útil obter a maior informação possível do pedido do utilizador, por outro lado não é viável, pelo tempo excessivo que pode decorrer, aplicar determinados métodos a todos os documentos ou frases considerados relevantes para o pedido.

Os métodos superficiais incluem:

- detecção de palavras-chave nos documentos para descobrir passagens promissoras;
- filtragem e classificação (*ranking)* destas passagens de acordo com o tipo de resposta esperado e a presença de determinados padrões sintácticos; eventualmente também por semelhança à pergunta ou co-ocorrência das palavras-chave;
- detecção de modelos (*templates*) predefinidos, na expectativa de a resposta ser uma reformulação ou inversão da pergunta – e.g. *Quem é X? X é Y.*

Para abordar os casos em que estas técnicas não são suficientes, têm sido aplicados métodos mais complexos, nomeadamente:

- reconhecimento das entidades nomeadas no texto;
- detecção de relações e co-referências entre as entidades;
- variações morfossintácticas;
- desambiguação lexical (*word-sense disambiguation*);
- inferência lógica (raciocínio abdutivo);
- outros tipos de raciocínio (senso comum, temporal, espacial, etc.)

2.3 Tipos de SRP

À medida que o campo de RP avança, os sistemas começam a partilhar determinadas características, mantendo a diferenciação em alguns aspectos. Em consequência, é actualmente possível classificar estes sistemas de acordo com características como:

a) recursos linguísticos utilizados;

b) técnicas de PLN e de processamento de documentos;

c) métodos de raciocínio;

d) capacidades de interacção com o utilizador;

e) tipos de perguntas a que conseguem dar resposta, dadas as capacidades do sistema.

Harabagiu e Moldovan [11] distinguem os sistemas em 5 classes:

(1) sistemas capazes de responder a perguntas factuais;

(2) sistemas com mecanismos de raciocínio simples;

(3) sistemas com capacidade de fundir informação de diferentes origens;

(4) sistemas com capacidade de raciocínio analógico;

(5) sistemas interactivos.

Classe 1 / sistemas que respondem a perguntas factuais:

Para estes sistemas, a resposta é um excerto de texto encontrado num só documento. A resolução do problema baseia-se em 2 processos essenciais:

(1) classificação da pergunta de acordo com o pronome interrogativo (e.g. *quem*) e a categoria das entidades nomeadas (e.g. *pessoa*, correspondendo à entidade *Sócrates*);

(2) extracção de respostas através de palavras-chave.

Outras técnicas de PLN utilizadas incluem a identificação de aposições[3] e paráfrases;[4]

3 Justaposição de sintagmas ou frases a outros sintagmas ou frases [20: pág.557]. No caso de sintagmas nominais, representa uma relação entre dois nomes: e.g. *"O João, um amigo nosso, (...)"*.

4 Diferentes formas de desenvolvimento de um texto ou de expressão do mesmo conceito. [13] E.g. *mergulhar no mar / furar as ondas*.

Classe 2 / sistemas com mecanismos de raciocínio simples:

A resposta é um excerto de texto, reconhecido como resposta através de inferências simples que a relacionam com a pergunta.

Recursos e técnicas de PLN:

- ontologias elaboradas que permitam organizar e aceder a conhecimento semântico;
- codificações de conhecimento pragmático e universal;
- reconhecimento de diferentes referências da mesma entidade (e.g. *UE* e *União Europeia*), e de referências semelhantes a entidades diferentes (e.g. *Paris* representando a capital francesa ou uma cidade do Texas, EUA);
- interpretação abdutiva de parágrafo e transformações metonímicas.[5]

Classe 3 / sistemas que fundem informação de várias origens:

Extraem informação parcial espalhada por diferentes documentos, e formulam uma resposta completa a partir deles. Os mais simples, que correspondem ao estado da arte actual, respondem a questões com resposta em lista. Outros sistemas mais complexos serão capazes de descrever sequências (s*cript-questions*) ou utilizar meios como a imagem e o vídeo.

A resolução de co-referências referida na classe 2 é importante nestes sistemas, devido à diversidade de fontes. Outras técnicas de PLN utilizadas:

- processamento semântico da pergunta mais avançado;
- sumarização automática.

Classe 4 / sistemas com capacidades de raciocínio analógico:

Respondem a questões especulativas cuja resposta poderá não estar explicitamente contida num documento. Decompõem a pergunta em vários processos de busca para extrair pistas que permitam formular a resposta.

Recursos:

5 Identificação de uma expressão a partir de outra, possibilitada por uma relação de contiguidade existente entre elas, que se exprime nas relações da causa pelo efeito, do todo pela parte, do continente pelo conteúdo, etc. E.g. *Beber um copo,* onde *copo* substitui *o conteúdo do copo*. [13]

- bases de conhecimento *ad hoc* geradas através de métodos de aprendizagem automática, tais como o agrupamento de documentos pelo tópico geral da pergunta;
- associadas a estas bases de conhecimento estão técnicas de raciocínio diversas: baseado em casos, baseado em evidências, temporal, espacial.

Classe 5 / sistemas interactivos:

Respondem a perguntas no contexto de interacções anteriores com o utilizador. Exigem diferentes algoritmos de resolução de referências, e taxonomias mais complexas de perguntas e respostas.

2.4 SRP baseados na Internet

Os motores de busca de documentos existentes *online* são ferramentas muito úteis pela sua funcionalidade e rapidez. No entanto, não preenchem todas as necessidades humanas no que diz respeito à obtenção de informação. Neste aspecto, o advento do paradigma *Web 2.0* foi um grande passo: propiciou a compilação e disponibilização colaborativa de informação em sítios como a Wikipedia, wikis especializadas e recentemente em páginas como o Yahoo! Answers que permitem aos utilizadores pedir informação a outros utilizadores – algo que já existia há algum tempo em domínios especializados como as Tecnologias da Informação, mas não estando geralmente acessíveis a utilizadores não registados.

Infelizmente, se os motores de busca tradicionais não se adequam a certos pedidos de informação específica, por sua vez as páginas de respostas colaborativas não permitem ao utilizador obter informação com a mesma rapidez. Existe um claro hiato na rede global entre os dois tipos de aplicação, que os SRP podem preencher cobrindo diversos casos de uso (como os referidos na Secção 2.2.1).

Os SRP independentes de domínio lidam com o mais variado tipo de questões. Portanto, em lugar de terem bases de conhecimento e ontologias específicas, têm de basear-se em ontologias gerais e conhecimento universal. Por outro lado, têm uma maior fonte de dados de onde podem extrair respostas. Um SRP baseado na Internet não se limita a proporcionar uma interface *online* para qualquer utilizador; usa a própria rede como fonte de documentos, aproveitando o permanente crescimento da informação disponível online. De facto, estes documentos

não se limitam a notícias, mas também a documentos provenientes de diversos nichos de interesses e conhecimentos.

A principal vantagem desta abordagem é o facto de existir uma grande variedade de textos e temas, propiciando a redundância. Numa colecção aberta e de grande dimensão, é provável que o mesmo assunto seja abordado de diferentes formas, o que facilita a tarefa do sistema pois este pode ser mais sensível a determinados padrões. Por outro lado, o processamento de conteúdo proveniente da Internet revela-se complicado pela variedade de origens, estilos ou mesmo erros, tanto ao nível da escrita como do próprio meio de disponibilização da informação (e.g. problemas de análise de código HTML ou de codificação de caracteres).

Do ponto de vista do utilizador da rede, este tipo de sistemas podem ser vistos como o próximo passo a seguir aos motores de busca. Especificamente, permitirão cada vez mais colmatar as deficiências dos serviços que já existem, tais como os referidos anteriormente:

- os tradicionais motores de busca – que não são adequados a certos pedidos de informação;
- as páginas onde outros utilizadores (especialistas ou não, consoante a política da página) respondem a perguntas previamente colocadas – que não garantem uma resposta em tempo útil.

A Figura 5 na página seguinte mostra um exemplo de uso do sistema AnswerBus, em Inglês. Apesar de ser possível obter a mesma informação através de um motor de busca de documentos, este sistema apresenta respostas imediatas, evitando que o utilizador tenha de seleccionar palavras-chave para efectuar o pedido ou percorrer páginas em busca da informação pretendida.

Actualmente, existe um interesse crescente em integrar a tecnologia de RP com a actual tecnologia de busca. Este cruzamento existe já há alguns anos no Ask (anteriormente conhecido por AskJeeves), um sistema que compara as perguntas efectuadas com as que tem na sua base de dados. No caso de não ser possível apresentar uma resposta, o programa devolve uma lista de documentos semelhante à doutros motores de busca mais populares. Segundo informação disponível na Wikipedia, estes motores de busca também começam a ser alvo de desenvolvimentos ao nível de RP [36].

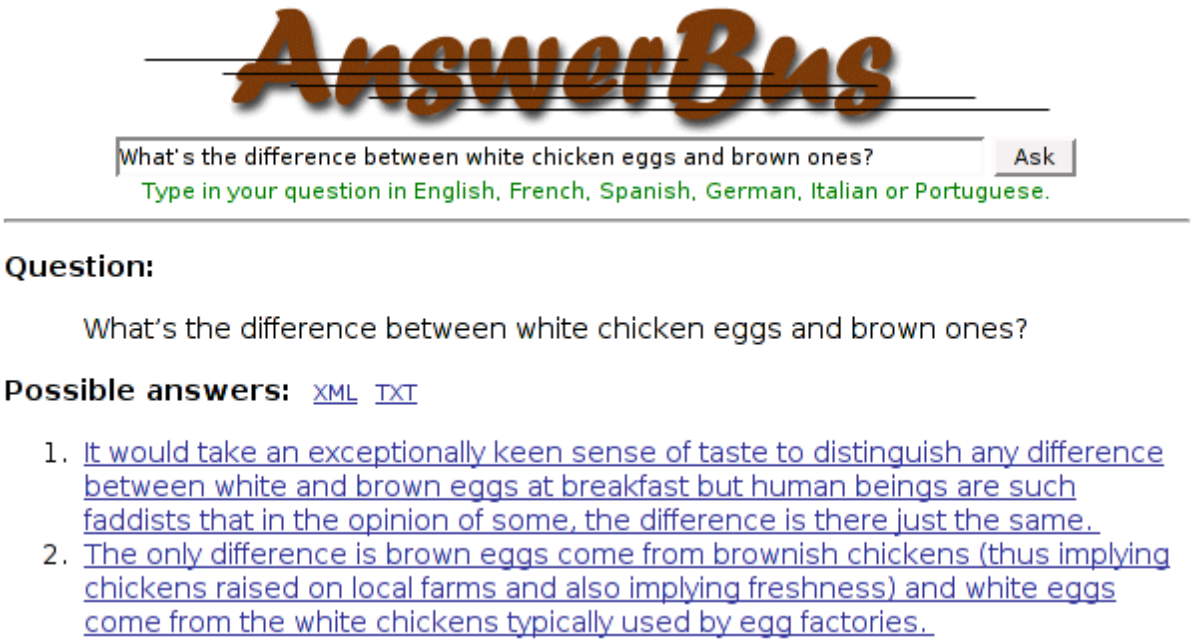

Figura 5: Pergunta ao SRP AnswerBus: "Qual é a diferença entre ovos de galinha brancos e castanhos?"

Independentemente destes esforços de integração, nos últimos anos foram já desenvolvidos alguns SRP baseados na Internet, sendo a maioria de âmbito académico. O START, do MIT, foi um dos primeiros sistemas a ter uma interface *web*, embora as suas fontes de respostas sejam variadas. Entre as referências do panorama actual, estão os sistemas AnswerBus, Brainboost e LCC. Uma lista dos SRP baseados na Internet, com os respectivos URL's, está disponível no Anexo A.

2.5 SRP em Língua Portuguesa

Como é habitual em diversos ramos de I&D, a língua inglesa tem recebido uma atenção privilegiada na área de RP. O sistema AnswerBus suporta línguas adicionais (Alemão e várias línguas românicas) através de métodos de tradução automática que convertem a pergunta para a língua inglesa e realizam todos os restantes passos nesta língua. Naturalmente, esta abordagem é bastante limitativa, pois não só adiciona um passo em que pode ocorrer perda de informação, como também não realiza busca e processamento específicos para as línguas em questão. A informação obtida é baseada apenas nos documentos existentes em Inglês e é devolvida ao utilizador nessa mesma língua, a não ser que se introduza mais um passo de tradução, onde

pode voltar a perder-se informação.

Dos poucos SRP em Português que estão em desenvolvimento com resultados publicados ou acessíveis destacam-se: o da empresa Priberam, o da Universidade de Évora e o Esfinge. Os dois primeiros estão vocacionados para operar sobre fontes estáticas de respostas, enquanto o último opera sobre os resultados de motores de busca de documentos na Internet.

Priberam

O sistema da Priberam [1, 2] destaca-se pelos resultados positivos que tem vindo a obter na conferência CLEF, nomeadamente 65% de acerto nas perguntas factuais testadas no evento de 2006 [20]. Actualmente o serviço de teste *online* encontra-se indisponível.

Este SRP tem uma abordagem semelhante à descrita na Secção 2.2.3, assentando em diversos recursos de PLN. Baseia-se ainda num pré-processamento dos documentos-fonte, com o objectivo de indexar informação semântica, domínios ontológicos e as categorias de perguntas às quais os documentos poderão responder. Pode realizar buscas locais num disco rígido ou na Internet, encontrando-se mais limitado neste último caso.

U. Évora

O sistema em desenvolvimento na Universidade de Évora [23], tal como o da Priberam, efectua um pré-processamento de documentos, mas com o objectivo de criar uma base de conhecimento. Posteriormente, efectua análise sintáctica baseada em gramáticas de restrições e análise semântica baseada numa teoria de representação do discurso [14]. Finalmente, realiza interpretação semântica/pragmática usando inferência lógica e ontologias.

Recentemente, estão em implementação métodos de aprendizagem automática nas tarefas de análise sintáctica/semântica (de documentos e pedidos) e de construção de ontologias. A diversidade de técnicas a testar levou a uma redefinição do sistema numa arquitectura multi-agente que permitirá conjugar os resultados dos diferentes métodos [25].

Este sistema encontra-se orientado a tipos de perguntas relevantes no domínio jurídico: lugares, datas, definições e outros casos específicos (e.g. *"Que crimes cometeu X?"*) [24].

Não houve participação na CLEF 2006; o resultado obtido em 2005 foi 22% de acerto nas perguntas factuais testadas.

Esfinge

O SRP Esfinge [8] evita o pré-processamento de documentos e aborda o problema de uma perspectiva estatística, baseando-se numa arquitectura proposta por Brill [5]. Procura aplicar técnicas simples a grandes quantidades de dados, explorando a redundância existente na Internet, onde o Português é uma das línguas mais utilizadas para além da língua inglesa [8]. Um passo fulcral no seu processo de resolução de pedidos consiste na detecção de n-gramas[6] nos excertos dos documentos fornecidos pelo próprio motor de busca utilizado (e não processando directamente os documentos listados pelo motor).

O melhor resultado deste sistema nas avaliações da CLEF de 2006 foi 24% de acerto nas perguntas factuais testadas. Está disponível um protótipo *online* no sítio da Linguateca, que procura devolver respostas exactas. Este protótipo apresenta ainda algumas limitações ao nível da relevância e do tempo de resposta.

6 Sequências de palavras de comprimento n.

Capítulo 3

O Sistema QueXting

O presente capítulo descreve pormenorizadamente o sistema, começando por uma panorâmica geral que distingue a abordagem de desenvolvimento, assim como os processos e recursos envolvidos na resolução de pedidos. Estes processos encontram-se agrupados em três módulos: processamento da questão, recolha de documentos e extracção de respostas.

3.1 Panorâmica

O QueXting é um SRP independente de domínio e dedicado a perguntas factuais, enquadrando-se no primeiro tipo de sistemas da classificação descrita no Capítulo 2 (Secção 2.3 – *Tipos de SRP*). Trata-se de um sistema baseado em texto não estruturado: a sua fonte de informação corresponde à colecção de documentos em língua portuguesa disponíveis a cada momento na Internet. É o primeiro sistema que processa directamente os documentos em língua portuguesa existentes na rede global.

3.1.1 Abordagem

A metodologia adoptada para um rápido desenvolvimento do sistema segue algumas directrizes estratégicas:

1. Adopção de uma arquitectura e de métodos independentes de idioma que têm sido testados e amadurecidos recentemente pela I&D em RP para outros idiomas.

2. Uso dos motores de busca disponíveis *online* para construir o corpus de documentos-fonte onde poderão ser encontradas respostas.

3. Exploração do potencial de recursos e ferramentas de PLN que o grupo NLX tem desenvolvido: analisador sintáctico, segmentador de frases, reconhecedor de entidades nomeadas, ontologias, etc. Estes componentes, específicos para o Português, permitem

adaptar o SRP a esta língua.

4. Desenvolvimento de ferramentas adicionais, específicas para RP em Português, nomeadamente um dicionário de perguntas/respostas.

5. Abordagem iterativa de desenvolvimento: após uma versão básica com os componentes essenciais, os módulos do sistema e a cobertura de perguntas são progressivamente desenvolvidos e optimizados com vista à obtenção de um desempenho que se aproxime do estado da arte em termos de resultados e tempo de execução.

A arquitectura referida atrás no ponto 1 foi ilustrada em traços gerais no Capítulo 2, Secção 2.2, sob a entrada *Processo de resolução*. Como referido nos relatórios das últimas conferências TREC, a arquitectura básica de SRP factuais não tem mudado, tendo amadurecido nos últimos anos [35]. Esta arquitectura geral e as ferramentas independentes de língua adoptadas em sistemas como o AnswerBus [39] constituem o ponto de partida que impulsionou o desenvolvimento inicial do QueXting.

Dadas as características próprias e as ferramentas aplicadas no QueXting, a arquitectura assume a forma do diagrama da Figura 6. (Os recursos entre parêntesis ainda não fazem parte da versão actual do sistema, mas as tarefas em que participarão são descritas adiante, em conjunto com as restantes tarefas.)

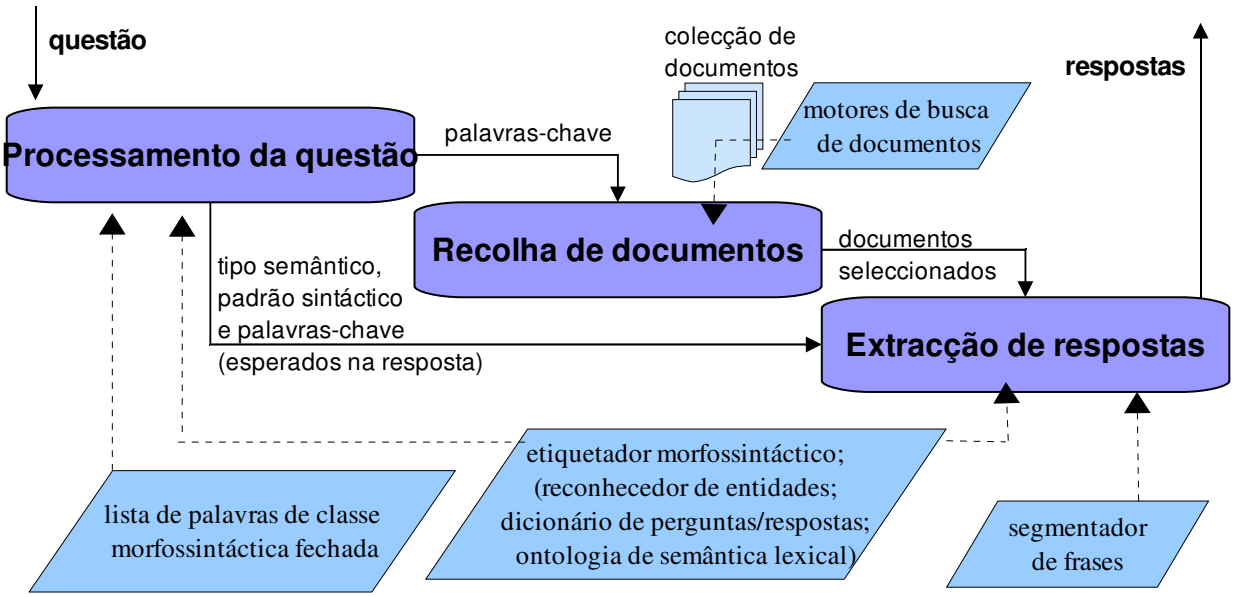

Figura 6: Processos e recursos do SRP QueXting

O sistema recebe pedidos de utilizadores através de uma página *web* onde a pergunta pode ser introduzida. O processo geral de recepção e resolução de pedidos de utilizadores é implementado no QueXting com a linguagem de programação Java (J2SE), efectuando-se através desta as chamadas necessárias às ferramentas linguísticas externas. A comunicação com a interface *web* é efectuada através de *sockets*, um processo que é facilitado pelo pacote *java.net*.

3.1.2 Processos e recursos

O processamento da questão envolve dois mecanismos essenciais: selecção de palavras-chave e identificação do tipo de pergunta e da resposta correspondente. A selecção de palavras-chave baseia-se numa lista de palavras a descartar e serve de base ao processo de identificação de documentos relevantes. A identificação do tipo de pergunta/resposta é realizada em primeira instância ao nível sintáctico, tendo especial relevância a presença de um pronome interrogativo ou de um sintagma nominal representando uma determinada entidade.

Neste momento, o sistema apenas realiza o processamento a este nível (sintáctico) pelo que não dispõe de informação sobre o tipo dessa entidade. O processamento da questão entrará no domínio semântico se for usado um reconhecedor de entidades nomeadas que distinga o tipo dessa entidade. Neste caso pode ser útil uma ontologia de semântica lexical que possibilite raciocínios sobre as entidades presentes na frase, permitindo aumentar a cobertura do reconhecedor ou mapear as entidades reconhecidas numa taxonomia de respostas própria do SRP.

Os documentos-fonte seleccionados pelos motores de busca são extraídos da *web* e processados de forma a obter os textos de onde poderão ser extraídas respostas concretas. As frases contidas nos textos são identificadas e avaliadas com base em critérios antecipados durante o processamento da questão (as próprias palavras-chave, o tipo semântico e o padrão sintáctico).

As respostas correspondem, portanto, a frases, cada uma delas extraída de um documento considerado relevante para a pergunta efectuada. Contudo, pode haver lugar à identificação e apresentação de respostas curtas (exactas) contidas nessas frases. As frases às quais se atribuiu maior valor e as respectivas expressões curtas são devolvidas ao utilizador através da página *web*, constituindo a resposta do sistema à pergunta introduzida.

Os aspectos principais da implementação destes processos em Java (e dos recursos externos que já se encontram aplicados) podem ser visualizados no diagrama da Figura 7:

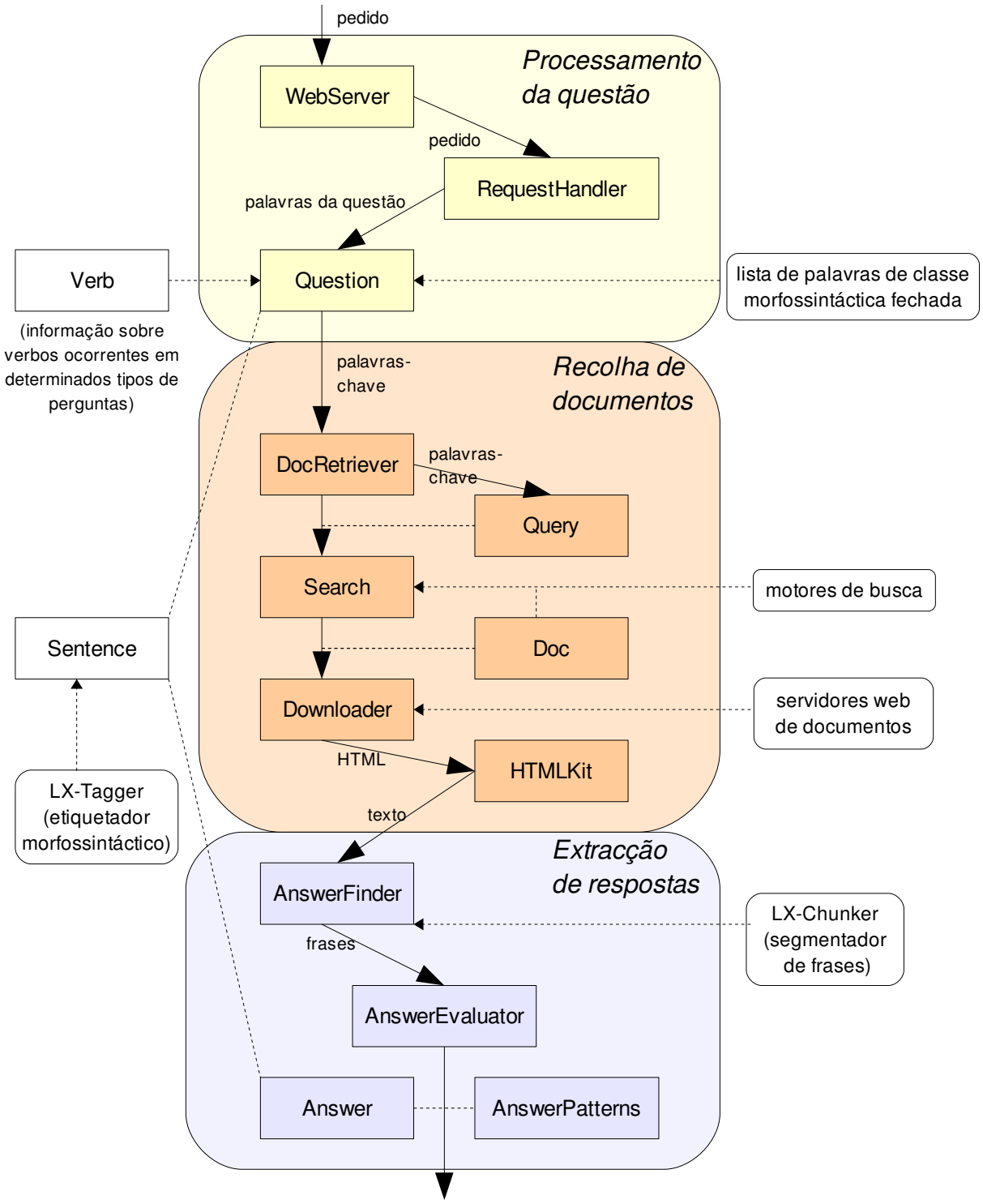

Figura 7: Processos e recursos actuais ao nível da implementação Java

Como exemplo da capacidade actual do sistema de detecção de respostas, o quadro seguinte apresenta a primeira resposta do QueXting à pergunta *"Quem foi a primeira pessoa a pisar na Lua?"*:

```
----------------------
RESPOSTA #1 (valor 5)
----------------------

Resposta exacta: Neil Armstrong
Frase completa: Neil Armstrong, primeira pessoa a pisar na Lua.
```

3.1.3 Interface de teste

Existem actualmente duas páginas *web* criadas no âmbito do projecto QueXting. A página de descrição do projecto[7] disponibiliza na rede um resumo sobre o QueXting, incluindo a sua motivação e os seus objectivos. O logótipo criado para esta página poderá ser adoptado também para a interface com o utilizador:

Figura 8: Logótipo

A página de interface com o sistema consiste numa versão interna de testes. Permite introduzir a pergunta, seleccionar os motores de busca de documentos e visualizar informação relativa ao funcionamento dos 3 módulos do sistema. A execução pode ser regular ou de teste, servindo este último caso para efectuar as buscas numa colecção estática de documentos que será abordada no Capítulo 4 (*Avaliação*). Um exemplo completo de execução do sistema para a pergunta da Figura 9 está disponível no Anexo C.

Figura 9: Envio de um pedido ao sistema

7 http://quexting.di.fc.ul.pt/

A caixa seguinte mostra o segmento inicial do resultado do pedido efectuado na Figura 9 anterior. Neste segmento inicial não é visível a resposta que será devolvida ao utilizador, mas apenas alguma informação de apoio ao desenvolvimento relativa às fases iniciais de resolução do pedido (processamento da questão e selecção de documentos). São apresentados dados relativos ao tipo de questão (tal como identificado pelo sistema), aos pedidos efectuados aos motores de busca (com as palavras-chave seleccionadas durante o processamento da questão) e ao primeiro documento devolvido por um dos motores:

```
Questão: "quem é edward norton?"
Tipo de questão: parcial "quem", predicativa
Processamento da questão em: 415 ms.


Query (é+edward+norton) enviada para o Google
URL a enviar: http://www.google.pt/search?q=é+edward+norton&lr=lang_pt&hl=pt-PT

Query (é+edward+norton) enviada para o ASKJ
URL a enviar: http://www.ask.com/web?q=é+edward+norton&dm=adv&advl=pt

Query (é+edward+norton) enviada para o Yahoo
URL a enviar: http://search.yahoo.com/search?p=é+edward+norton&fl=1&vl=lang_pt

Query (é+edward+norton) enviada para o MSN
URL a enviar: http://search.msn.com/results.aspx?q=é+edward+norton+language%3Apt


URL:www.geocities.com/Hollywood/Theater/3451/spotlight.html
URL ORIGINAL: www.geocities.com/Hollywood/Theater/3451/spotlight.html
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 1
TIPO: HTML
SNIPPET: Gostoso! - e adjetivos do tipo - tenha certeza de apenas uma coisa: este indivíduo
não é Edward Norton Colocando as cartas na mesa, ele não tem nada demais. É narigudo, tem
uns dentes de ...
```

Numa execução normal do sistema são apresentados vários documentos e para cada um deles é apresentada a informação proporcionada pelos motores de busca, tal como o endereço, o tamanho e alguns excertos de texto.

O quadro seguinte apresenta a primeira resposta do sistema à pergunta da Figura 9, acrescida de informação sobre o documento e o tempo de processamento. Antes da resposta propriamente dita é possível ver o excerto final do descarregamento de documentos e a contagem de frases candidatas a resposta (i.e. contendo um número mínimo de palavras-chave):

```
* Download de www1.folha.uol.com.br/folha/ilustrada/ult90u327850.shtml concluido.

* Download de cineminha.uol.com.br/materia.cfm?id=5049 concluido.

Frases candidatas: 90
Frases inuteis: 1082
```

```
----------------------
RESPOSTA #1 (valor 6)
----------------------

Resposta exacta: ator perfeito
Resposta exacta: cristalina
Frase completa: Verdade pura e cristalina, Edward Norton é um ator perfeito e não está para
brincadeira.

Data: Mon Nov 15 19:55:37 WET 1999
Documento: www.geocities.com/Hollywood/Theater/3451/spotlight.html
Motor: MSN (rank 1)

Busca e parse de resultado: 514 ms.
Download e parse de documento: 492 ms.
Extraccao de resposta: 258 ms.
Tempo total: 1679 ms.
```

As secções seguintes deste capítulo descrevem em pormenor cada um dos passos de execução do sistema, clarificando também o papel dos recursos utilizados.

3.2 Processamento de questões

Para um ser humano, uma frase adquire um significado global a partir do significado das expressões que a compõem e das regras implícitas que gerem a sua combinação. Esta perspectiva, denominada *princípio da composicionalidade*, apesar de não totalmente abrangente,[8] pode ser aplicada no processamento computacional de frases.

Na prática, a importância de cada palavra para o processamento da pergunta é variável; existem claros padrões de palavras e construções sintácticas nas perguntas que são os pilares fundamentais para obter o seu significado. Uma aplicação automática de RP pode ter a capacidade de compreender alguns aspectos essenciais da pergunta, através da identificação dos seguintes aspectos:

a) palavras-chave da pergunta que deverão estar presentes na resposta;

b) palavras que constituem pistas sobre o tipo de pergunta/resposta;

c) padrões sintácticos esperados na resposta.

3.2.1 Palavras-chave

A identificação de palavras-chave na pergunta tem duas vantagens fundamentais: optimiza a selecção de documentos e permite uma rápida pré-selecção das frases contidas nesses docu-

8 Não cobre casos que exigem um contexto alargado, como por exemplo o uso de metáforas, metonímia, ironia.

mentos. A optimização da selecção de documentos é conseguida por afinação do pedido (*query*) que é enviado aos motores de busca para encontrar documentos relevantes. Para além da identificação das palavras da pergunta que devem fazer parte da *query*, a sua criação pode envolver processos de substituição, simplificação ou expansão.

O primeiro destes processos corresponde ao uso de lemas, ou seja de formas morfologicamente normalizadas das palavras originais (e.g. no caso de um verbo: uso do seu infinitivo). No processo de simplificação, as palavras funcionais (preposições, conjunções, etc.) são eliminadas; as palavras não funcionais muito frequentes também poderão sofrer o mesmo destino em virtude de terem um impacto negativo na precisão e no desempenho do sistema. O processo de expansão consiste no uso de sinónimos ou termos semanticamente relacionados com os existentes na pergunta, com o objectivo de aumentar a cobertura.

Actualmente, o sistema QueXting realiza um destes três processos: a simplificação da *query*. A sua implementação baseia-se num léxico de palavras de classes morfossintácticas fechadas: todas as palavras da pergunta que façam parte desse léxico são automaticamente eliminadas. Normalmente, os motores de busca realizam, eles próprios, uma filtragem deste tipo; no entanto, a sua realização pelo próprio sistema permitirá, mais tarde, avaliar as possíveis respostas em termos das palavras-chave mais relevantes.

3.2.2 Tipo semântico e padrões sintácticos

Certas palavras, além de serem importantes para efeitos de identificação de documentos-fonte ou pré-selecção de frases candidatas, constituem pistas com valor semântico particularmente significativo. A identificação destas pistas semânticas na pergunta permite distinguir de forma mais informada as respostas adequadas.

Por exemplo, à pergunta *"Quem foi a segunda pessoa a pisar a Lua?"* corresponderá certamente uma resposta cuja expressão denota uma pessoa. O mesmo se prevê para a pergunta *"Qual foi o segundo astronauta a pisar a Lua?"* embora através dum raciocínio diferente. Como se extraem estas conclusões? No primeiro caso, as palavras *quem* e *pessoa* constituem evidência suficiente; contudo, o segundo caso exige o conhecimento ou raciocínio de que um astronauta é uma pessoa.

Como se pode deduzir dos exemplos anteriores, um factor que contribui para a extracção

de significado da pergunta é a presença de um pronome interrogativo, como por exemplo: *quem*, *quando*, *onde*, *quanto*, *qual* ou *que.* Para muitas perguntas, este factor pode ser suficientemente discriminativo para determinar o tipo de resposta a esperar: a maioria dos pronomes deste tipo define imediatamente (embora nem sempre da forma mais específica) o tipo de conceito ou entidade sobre o qual a pergunta é feita e a resposta é esperada.

Esse não é o caso dos dois últimos pronomes, que se aplicam a diversos tipos de perguntas (diferentes tipos de entidades ou eventos). Uma pergunta do tipo *"Qual foi X?"* exige já algum processamento semântico para que se possa atribuir um tipo de resposta. Por exemplo, a pergunta *"Qual foi o membro fundador de Y?"* deverá ter como resposta uma entidade do tipo *pessoa*; no entanto a pergunta *"Qual foi o dia em que ocorreu Z?"* exige como resposta uma data.

Nestes casos, para encontrar o tipo de entidade que deverá estar contida na resposta, será necessário usar um **reconhecedor de entidades nomeadas**. Uma ferramenta deste tipo está em desenvolvimento no grupo NLX e planeia-se a sua aplicação ao QueXting. Os tipos de entidades reconhecidas por esta ferramenta constituem a primeira base de tipos semânticos de respostas que poderão ser dadas pelo sistema.

Contudo, poderá ser definida uma taxonomia de respostas específica para a tarefa de RP, que sirva de base a um **dicionário de perguntas/respostas**. Este dicionário associaria às classes de perguntas as classes de respostas esperadas. As classes de respostas reconhecidas (*pessoa*, *data*, *local*, etc.) podem estar hierarquizadas em sucessivos níveis de especificidade. Cada tipo de resposta teria as suas palavras-chave e padrões sintácticos esperados, nos quais se basearia a avaliação das potenciais respostas.

Usando um dicionário deste tipo, o reconhecimento de entidades seria complementado por uma **ontologia de semântica lexical**, semelhante à WordNet [9] que foi criada para a língua inglesa. Uma ontologia deste tipo permite efectuar raciocínios tais como "*astronauta* é uma entidade mais específica do tipo *pessoa*", o que ajuda a identificar os tipos de entidades que ocorrem nas frases (na pergunta e nas potenciais respostas).

Dado que os tipos semânticos de entidades e respostas ainda não são reconhecidos pelo QueXting, este é orientado por uma **taxonomia de perguntas**: foi efectuado um levantamento

de vários tipos de perguntas factuais possíveis e um tipo de pergunta foi escolhido como primeiro exercício de desenvolvimento. Procedeu-se, assim, ao desenvolvimento de capacidades de detecção e processamento de perguntas parciais, predicativas, do tipo *"Quem...?"*.

Perguntas *"Quem..."* predicativas

Eis um exemplo deste tipo de pergunta:

Quem	*é*	*o novo presidente da câmara municipal de Lisboa*	*?*
pronome interrogativo	**verbo**	**sintagma nominal**	

O reconhecimento de uma pergunta deste tipo baseia-se na identificação do pronome interrogativo *Quem*, no princípio da pergunta, possivelmente seguido de uma expressão dita de clivagem[9] (*"é que"*) e de uma forma verbal predicativa.[10] Após o verbo, é esperado o sintagma nominal (SN) sobre o qual incide a pergunta.

Em termos práticos, o reconhecimento é conseguido usando uma expressão regular que denota este tipo de pergunta, com os componentes referidos no parágrafo anterior, através das funcionalidades do pacote *java.util.regex*.

Não é feita uma verificação de que o último constituinte se trata realmente de um SN. Essa acção impediria o reconhecimento de algumas perguntas válidas pois as capacidades de detecção de SN's do QueXting ainda não abrangem os SN's que contêm conjunções ou sintagmas preposicionais.

Os padrões sintácticos esperados em respostas a este tipo de perguntas foram estudados e definidos, sendo usados pelo sistema para identificar respostas exactas. Pormenores adicionais sobre este ponto são referidos adiante na Secção 3.4 *(Extracção de respostas)*.

3.3 Recolha de documentos-fonte

A obtenção de fontes de respostas, por ser baseada na Internet, não se limita apenas a um processo de selecção. Obriga a um processo mais complexo que envolve o descarregamento de documentos e uma análise desses documentos que permita extrair texto útil, evitando duplica-

9 Para uma definição completa de expressões de clivagem, cf. [21], pp. 685-692.
10 Verbos predicativos contemplados: *ser*, *estar*, *ficar*, *continuar*, *permanecer*, *tornar-se*.

dos tanto quanto possível. Embora a redundância de informação possa ser útil se ocorrer em diferentes contextos ou abordagens, um documento cujo texto útil é exactamente igual a outro já processado deve ser descartado, por razões de eficiência e variedade.

3.3.1 Identificação

A recolha de documentos relevantes é feita usando motores de busca existentes *online,* a partir das palavras-chave seleccionadas pelo módulo de processamento de questões. Os motores que o sistema foi preparado para usar incluem o *Yahoo!*, *Google*, *Live (MSN)* e *Ask*.

A comunicação com os motores de busca é conseguida através das funcionalidades do pacote *java.net*. O sistema envia para os motores um pedido com as palavras-chave (embebidas no URL que é usado para estabelecer ligação com cada motor), recebe as páginas de resultados e identifica no código HTML das páginas de resultados os dez primeiros documentos devolvidos por cada motor, evitando URL's repetidos.[11]

É, assim, criada uma colecção de textos que servirá de fonte para procurar as respostas específicas, constituída pelos documentos considerados mais relevantes pelos motores de busca.

O sistema está preparado para processar documentos HTML. Contudo, é de notar que a Internet é constituída por documentos de vários formatos. Alguns motores de busca disponibilizam versões HTML de documentos originalmente noutros formatos (e.g. PDF, DOC, PPT). Esta potencialidade é explorada, quando presente nas páginas de resultados dos motores, e permite ao sistema poupar tempo ao lidar com diferentes formatos sem precisar de implementar nem executar um conversor ou *parser* para cada tipo de documento.

3.3.2 Recolha

O descarregamento dos documentos seleccionados é efectuada automaticamente através das funcionalidades do pacote *java.net*, estabelecendo ligações com os servidores onde os documentos se encontram alojados e recebendo as páginas HTML das quais pode ser extraído o texto útil. No descarregamento e salvaguarda dos documentos surge outro desafio provocado pela utilização da Internet como fonte de respostas: como interpretar correctamente os caracteres que compõem os textos?

11 Estas funcionalidades baseiam-se num esboço de código experimental e parcelar desenvolvido no NLX por Luís Aguilar e corresponderam ao ponto de partida do desenvolvimento do QueXting.

Um documento de texto não fica totalmente definido pelos códigos dos caracteres que o compõem; ao ser criado, o documento é salvaguardado com um determinado esquema de codificação de caracteres (*character encoding*). Cada esquema suporta um conjunto diferente de caracteres, denominado *charset (character set)*, o que significa que pode acontecer que os mesmos códigos correspondam a diferentes caracteres em diferentes *charsets*. Isto é especialmente visível nos caracteres acentuados ou simbólicos. Portanto, ao descarregar um documento da rede, torna-se necessário reconhecer primeiro o esquema de codificação utilizado.

Existem duas possibilidades para detectar esta informação, e ambas são usadas pelo QueXting:

(a) Logo após estabelecer ligação com o servidor remoto que aloja o documento, pedir o *content-type* desse documento. Este procedimento é conseguido através do método:

- `java.net.HttpURLConnection.getContentType()`

(b) Após o descarregamento, detectar se existe uma entrada no cabeçalho do código HTML que defina o esquema de codificação que deve ser utilizado para ler a página correctamente. Este procedimento é efectuado pelo analisador de HTML descrito na secção seguinte, sendo apenas necessário dar-lhe a indicação de que deve estar atento às declarações de *charset.*

Por omissão, os caracteres são interpretados com o esquema `Windows-1252`, que inclui todos os caracteres válidos em HTML de outro esquema usado frequentemente, o `ISO-8859-1`. Esta opção foi tomada devido ao facto de alguns documentos serem erroneamente marcados com o esquema `ISO-8859-1`, quando na prática usam caracteres simbólicos (e.g. aspas curvas e travessões longos) que não fazem parte dele – o que leva a uma falha de reconhecimento desses caracteres e à sua substituição por pontos de interrogação.

No entanto, nos casos em que é detectada uma referência ao esquema `UTF-8`, este será o utilizado para interpretar o documento. Os três esquemas referidos ao longo desta secção são aqueles que são suportados neste momento pelo QueXting e, até aqui, têm coberto todos os documentos processados pelo sistema.

Na prática, o QueXting realiza sempre o descarregamento sob a forma de octetos (*bytes*), até ao tamanho máximo de 512 KB definido como parâmetro do sistema, para evitar documentos demasiado longos. Apenas na fase seguinte é efectuado o reconhecimento dos caracte-

res que estão contidos nesses octetos.

3.3.3 Análise

Cada documento extraído da rede é sujeito a análise (*parsing*), com o objectivo de reconhecer e extrair o texto útil, i.e. aquele que não corresponde ao código de apresentação da página. É usado o analisador de HTML existente na biblioteca *swing* do Java para este processo. Este analisador detecta no código HTML as indicações de esquema de codificação de caracteres referidas na alínea (b) da secção anterior.

Caso essas indicações não existam, o analisador pode usar o *charset* eventualmente encontrado através do método da alínea (a) acima. Para tal, a chamada ao analisador deve ser feita com um objecto `Reader` que tenha esse *charset* definido para a leitura. Eis a assinatura do método que efectua a chamada ao analisador (notar o objecto Reader referido no primeiro parâmetro; o terceiro parâmetro que diz respeito à detecção de *charset* no código HTML):

- ```
  java.swing.text.html.HTMLEditorKit.getParser().parse(
    Reader r,
    ParserCallback cb,
    boolean ignoreCharSet )
  ```

Para utilização deste analisador é necessário definir os métodos que serão chamados quando o analisador encontra elementos de texto, comentários ou determinada etiqueta HTML. Estes métodos podem ser definidos estendendo a classe `HTMLEditorKit.ParserCallback` (um objecto deste tipo deverá ser o segundo argumento do método acima). A título de exemplo, as linhas seguintes representam a implementação do método que lida com elementos de texto:

```
public synchronized void handleText (char[] data, int pos)
{
    // skipping style+script code that is mistaken for text
    if (!skipCode)
    {
        if (data[0]=='>' || data[0]=='<') {
        // skipping leftover shifts from certain html tags
            for(int i=1; i < data.length; i++)
                text.append(data[i]);
            text.append(' ');
        } else
            text.append(data).append(' ');
    }
}
```
```

Com alguma frequência, há código de *scripting* ou de estilo CSS (Cascade Style Sheets) que é interpretado como texto pelo analisador de HTML, o que resulta em texto com algum "lixo". O analisador foi parametrizado para ignorar tudo o que está contido entre as etiquetas `<style>` e `</style>`, ou `<script>` e `</script>`. Infelizmente, o próprio analisador parece ter limitações pois há etiquetas desse tipo que não são reconhecidas, o que implica que algum desse lixo ainda se mantém nos textos. Apesar de grande parte dele já ser evitado, poderá ser ponderado o uso de analisadores alternativos.[12]

Além deste problema, nem todo o texto real corresponde efectivamente a conteúdo útil para o sistema, dado que é quase impossível distinguir menus e cabeçalhos do conteúdo propriamente dito. Uma medida que foi implementada com sucesso foi o descarte das frases que contêm o caracter '|', que surge frequentemente em listas de opções de navegação. Esta medida é tomada no passo seguinte do QueXting, i.e. no módulo de extracção de respostas, já que é este que processa o texto obtido de forma a obter frases individuais.

3.3.4 Evitando duplicados

Para evitar documentos repetidos, é realizada uma verificação dos URL's referenciados pelos diferentes motores de busca: URL's iguais são processados apenas uma vez. No entanto, podem existir na rede documentos de conteúdo idêntico, alojados em URL's diferentes. Uma funcionalidade adicionada a este módulo do sistema foi a detecção de documentos redundantes em termos de conteúdo.

As páginas podem ter cabeçalhos e menus diferentes, o que elimina opções de comparação directa. Logo, como ponto de partida, foi criada uma função que verifica se os excertos de um documento que foram seleccionados pelo motor de busca e apresentados nos resultados fazem todos parte do conteúdo do outro documento (e vice-versa).

Por exemplo, para um determinado documento, um motor de busca fornece os seguintes excertos, originalmente separados por reticências (e um espaço de cada lado destas):

```
Edward Norton (Boston, 18 de Agosto de 1969) é um ator norte-
americano. ... O Wikiquote tem uma coleção de citações de ou so-
bre: Edward Norton. ...
```

12 Este ponto será novamente abordado na secção *Trabalho futuro* do capítulo final.

A verificação consiste em comparar os dois excertos deste documento com os excertos dos documentos anteriormente processados.

Esta forma de comparação dos documentos é bastante simples, o que poderá constituir uma desvantagem – ocasionalmente podem ocorrer falsos positivos, i.e. documentos erroneamente considerados redundantes devido a um deles conter os excertos do *snippet* do outro documento – mas também uma vantagem em termos de velocidade de processamento. Numa fase de optimização do sistema, métodos alternativos de comparação de documentos deverão ser pesquisados e testados.

3.4 Extracção de respostas

Obtidos os documentos-fonte, resta ao sistema encontrar excertos que contenham uma resposta específica. Esperam-se respostas curtas e directas (além de relevantes), já que o sistema se destina a responder a perguntas factuais.

Antes de mais, as frases têm que ser identificadas, ou seja, isoladas das restantes frases do documento para processamento individual. Depois podem ser avaliadas à luz de diversos factores como o tipo semântico, o padrão sintáctico e a presença das palavras-chave. As eventuais semelhanças são pesadas por uma função que determina o valor da frase; as frases de maior valor serão as apresentadas ao utilizador.

Um factor adicional de desempate que está a ser usado na avaliação de respostas consiste em privilegiar, entre frases do mesmo valor, aquelas que são oriundas de documentos mais recentes.

3.4.1 Segmentação em frases

O módulo de extracção de respostas deve, em primeiro lugar, extrair excertos dos documentos recolhidos. Para este fim é usado o segmentador LX-Chunker, desenvolvido anteriormente pelo grupo NLX, aplicando-o sobre o texto de cada um dos documentos recolhidos.

A ferramenta permite marcar os excertos (parágrafos e frases) que compõem o texto. A sua funcionalidade consiste em tomar o texto que lhe é dado como entrada e assinalar: o início e fim de parágrafos, respectivamente com as marcas `<p>` e `</p>`; e o início e fim de frases,

respectivamente com as marcas <s> e </s>. O seu método de funcionamento é baseado num autómato finito que representa diferentes estados do discurso.[13] As linhas da caixa seguinte apresentam um exemplo de aplicação, com o texto original à esquerda e o resultante à direita:

O ministro das Finanças, António Sousa Franco, garantiu, ontem, no Parlamento, que a execução orçamental deste ano, em matéria de receitas e em matéria de despesa, está a decorrer acima das previsões mais optimistas. Sousa Franco admitiu mesmo que, a manter-se neste ritmo, o défice orçamental poderá vir a situar-se abaixo do inicialmente previsto.	** <s>**O ministro das Finanças, António Sousa Franco, garantiu, ontem, no Parlamento, que a execução orçamental deste ano, em matéria de receitas e em matéria de despesa, está a decorrer acima das previsões mais optimistas. **</s><s>**Sousa Franco admitiu mesmo que, a manter-se neste ritmo, o défice orçamental poderá vir a situar-se abaixo do inicialmente previsto. **</s> **

Como referido anteriormente, uma característica inevitável e problemática nos documentos extraídos da *web* é a dificuldade em diferenciar o conteúdo útil, que constitui realmente o texto do documento. Esta questão já havia levantado o problema da inclusão de código (*scripts* e CSS) no texto dos documentos extraídos; no contexto da segmentação em frases surge um novo problema.

Numa página HTML, tudo o que não é código de HTML ou das suas extensões, é considerado texto, pelo que expressões que façam parte de menus ou títulos serão incluídas no texto do documento. No entanto, o segmentador não estava preparado para este tipo de textos, não conseguindo identificar frases que não terminam com sinais de pontuação, a não ser que a frase termine com um caracter de fim-de-linha.

Infelizmente, o módulo de recolha de documentos extrai o texto dos documentos usando um analisador de HTML que ignora estes caracteres de fim-de-linha, resultando em texto que fica todo numa só linha. A solução consiste em editar o módulo de recolha de documentos, estendendo o analisador para reconhecer e lidar com as terminações de linha. O analisador passou a ter uma regra para, quando encontrar uma etiqueta HTML do tipo e
, acrescentar ao texto do documento um caracter de fim-de-linha.

Qualquer caracter de fim-de-linha que escape ao LX-Chunker é detectado após a chamada ao segmentador, através do próprio código Java, para separar as expressões à esquerda e direi-

13 Para mais pormenores, consultar [3, 27].

ta desse caracter em frases independentes, evitando problemas curiosos como este exemplo real apresentado como resposta pelo QueXting:

> *Ex-presidente da Nicarágua é mantido em prisão de alta segurança*
> *da France Presse , em Manágua*

A primeira linha da resposta representa um título de uma notícia e a segunda linha representa a proveniência dela – nunca deveriam ser apresentados em conjunto, pois fora do contexto da página *web* perde-se a percepção do que representam. Assim, são separadas pelo código Java em frases distintas.

As linhas de código seguintes mostram como a ferramenta LX-Chunker pode ser chamada a partir do Java. Neste código é usada uma *thread* separada (`WriteToChunker`) para passar o texto do documento ao Chunker, de forma a impedir que o preenchimento do tampão de entrada (*input buffer*) desta ferramenta bloqueie o programa principal. A instrução `evalSentence` é uma chamada à função de avaliação baseada em palavras-chave, descrita na secção seguinte.

```
// launching the LX-Chunker to segment the Doc's text into sentences
Runtime rt = Runtime.getRuntime();
Process proc = rt.exec( Prefs.LXCHUNKER_PATH );

// providing text to chunker
// (in separate thread to avoid blocking due to full input buffer)
BufferedWriter bw = new BufferedWriter( new OutputStreamWriter(
proc.getOutputStream(), "UTF-8" ));

new WriteToChunker(bw).start();

// reading the resulting sentences
BufferedReader br = new BufferedReader( new InputStreamReader(
proc.getInputStream(), "UTF-8" ));

Scanner chunkOut = new Scanner(br).useDelimiter("<s>");
while (chunkOut.hasNext())
{
  String sentence = chunkOut.next();

  // removing chunker tags and trailing newlines...
  if(sentence.endsWith("<p>"))
    sentence = sentence.substring(0, sentence.length()-3);
  if(sentence.endsWith("</p>"))
    sentence = sentence.substring(0, sentence.length()-4);
  if(sentence.endsWith("</s>"))
    sentence = sentence.substring(0, sentence.length()-4);
  while(sentence.endsWith("\n"))
    sentence = sentence.substring(0, sentence.length()-1);
```

```
  // evaluating each sentence and store candidate answers
  if (sentence.contains("\n")){
    // deals with lx-chunker undetected newlines
    String[] sentences = sentence.split("\\n");
    for (String s : sentences)
      if (!s.isEmpty())
        evalSentence(s);
  } else
      evalSentence(sentence);

}//endWhile

// closing the Scanner and the LX-Chunker subprocess
chunkOut.close();
int exitVal = proc.waitFor();
if(exitVal != 0)
  Util.report( null, true,
               "Chunker Error?\nProcess exit value: " + exitVal );
```

As frases encontradas nos documentos são alvo de uma avaliação que permite distinguir as melhores candidatas a fazer parte da lista final de respostas. As características das frases (em termos do tipo de resposta que constituem, dos seus padrões sintácticos e da presença das palavras-chave) são comparadas com as características esperadas, determinadas durante o processamento da questão.

3.4.2 Detecção de palavras-chave

Na prática, os factores de avaliação não são verificados num só passo pela mesma função de avaliação. A primeira função de avaliação aplicada permite realizar uma primeira filtragem rápida e baseia-se na função utilizada no AnswerBus [39]. Esta heurística mede o número de palavras-chave, exigindo um número mínimo de palavras-chave na frase relativamente ao número de palavras-chave identificadas na pergunta inicial. As frases em que esse limite mínimo não é atingido recebem o valor 0; as restantes recebem um valor tanto maior quanto maior for o número de palavras-chave efectivamente presentes.

Em suma, são seleccionadas como potenciais respostas as frases que obedecem à seguinte fórmula (onde **Q** corresponde ao número de palavras-chave da pergunta e **q** ao número de palavras encontradas na frase candidata):

$$q \geq \left\lfloor \sqrt{Q-1} \right\rfloor + 1$$

Por exemplo, para a pergunta *"Quem é Edward Norton?"* seriam consideradas palavras-chave as seguintes: *é, Edward* e *Norton.* A função de avaliação, que consiste no membro direito da inequação acima, resulta no valor 2. Portanto, qualquer frase contendo pelo menos duas das três palavras-chave é uma candidata a aparecer na lista final de respostas devolvida ao utilizador.

Foram implementadas duas formas de detecção de palavras-chave no texto dos documentos. A primeira realiza uma detecção simples e rápida, ignorando naturalmente as diferenças de caixa alta/baixa (*character case*). A segunda alternativa (de momento, não utilizada) foi criada para o caso de se querer detectar palavras muito semelhantes, ignorando diferenças mínimas adicionais como, por exemplo, ao nível da acentuação.

Esta função básica foi estendida para privilegiar frases em que as palavras-chave ocorrem conjuntamente. Assim, independentemente do número de palavras-chave detectadas na frase, se todas elas surgirem em sequência (em qualquer ordem), então a frase recebe um valor de pontuação adicional. Esta verificação é independente da ordem para cobrir os casos em que a estrutura sintáctica da resposta corresponde a uma inversão da estrutura sintáctica da pergunta. É concretizada através da geração rápida de permutações das palavras-chave.

Retomando o exemplo anterior, a frase *"Edward Norton é um actor norte-americano"* contém as três palavras-chave. Como todas elas ocorrem em sequência, a frase recebe um valor adicional de pontuação, apesar de as palavras surgirem numa ordem diferente daquela em que ocorrem na pergunta. A frase *"Edward Norton protagoniza Fight Club"* teria só duas palavras-chave, mas também receberia um valor adicional por ambas surgirem juntas.

Concluindo, a verificação de palavras-chave corresponde a um primeiro passo de filtragem. O número de respostas candidatas é limitado (a um valor que é parâmetro do sistema) pelo que, quando a lista está cheia, uma nova frase só entra na lista se tiver um valor maior que outra frase lá existente. Apenas após todas as frases passarem por esta verificação de palavras-chave é que os restantes factores de avaliação são aplicados. Este procedimento permite aplicar os métodos de avaliação mais custosos a um menor número de frases – apenas àquelas que foram suficientemente boas, em termos de palavras-chave, para ficar na lista final de respostas, cujo tamanho é também um parâmetro do sistema.

Note-se que, se forem encontradas poucas frases com as palavras-chave, todas elas pode-

rão fazer parte da lista final. Apenas nos casos em que existem muitas candidatas é que as menos valiosas em termos de palavras-chave ficam eliminadas à partida. Este processo contribui para um desempenho temporal mais uniforme: as perguntas que levam a muitas frases candidatas não têm um impacto demasiado grande no sistema.

3.4.3 Detecção de padrões sintácticos e do tipo semântico

As frases que passarem o filtro das palavras-chave e forem suficientemente boas (ou em número total pequeno) para ficar na lista final de respostas são novamente analisadas, desta vez em termos de padrão sintáctico e tipo semântico.

É aqui evitado o processamento e posterior apresentação de frases demasiado parecidas: são consideradas semelhantes as frases que partilham uma sequência de 80% das palavras da maior frase, sendo a menor frase descartada nesse caso. A razão da manutenção da maior frase prende-se com a existência de maior contexto para o utilizador; a maior extensão da frase poderá ser compensada com a identificação de expressões dentro dessa frase que constituam respostas exactas.

Vejamos um exemplo plausível, com duas frases que seriam consideradas semelhantes, sendo descartada a segunda:

- *Jorge Sampaio tornou-se o presidente da República Portuguesa em 1996.*
- *Jorge Sampaio tornou-se o presidente da República Portuguesa.*

A primeira frase providencia claramente mais contexto (neste caso, de carácter temporal), no entanto o seu maior comprimento pode ser compensado pela identificação de expressões curtas (dependentes da pergunta) como *Jorge Sampaio*.

Foi implementada uma classe Java denominada `AnswerPatterns` que compila, através de expressões regulares, os padrões sintácticos esperados para os tipos de perguntas. Por enquanto, são cobertas as perguntas parciais predicativas do tipo **"Quem...?"**. Esta classe irá cobrir mais casos de perguntas, à medida que for posta em prática a sua cobertura; poderá inclusivamente ser adaptada para compilar os padrões por tipo de resposta (e.g. *pessoa*, *organização*, etc. – de acordo com a taxonomia de respostas), em vez de tipos de perguntas.

Para o tipo de pergunta referido, já estão definidos os padrões sintácticos esperados. Estes

são procurados nas frases candidatas e valorizam a frase em 1 ponto adicional, além de conterem uma resposta exacta – i.e. uma expressão que deve responder de forma clara e concisa à pergunta do utilizador, consistindo numa entidade de determinado tipo semântico, que pode ser comparada com o tipo esperado se este for detectado por um reconhecedor de entidades nomeadas.

Como referido na Secção 3.2, está planeado para breve o uso duma ferramenta desse tipo, que permitirá processar as frases candidatas para detectar entidades e atribuir-lhes um tipo semântico. A distância semântica entre o tipo atribuído e o tipo antecipado no processamento da questão será um dos factores de *ranking* da resposta.

Com a sua capacidade de detectar entidades específicas, espera-se que esta ferramenta evitar problemas como o deste exemplo real, proveniente dos resultados da primeira conferência CLEF (2003) e traduzido para Português:

- Pergunta: *"Quem inventou o rádio?"*
- Resposta: "*Alguém"*

Respostas a perguntas "Quem..." predicativas

No coração dos padrões sintácticos de resposta para este tipo de pergunta, está a entidade-alvo sobre a qual é feita a pergunta. Dado o sintagma nominal (SN) extraído da pergunta, outro SN contendo uma resposta poderá ocorrer à esquerda ou à direita do primeiro.

Exemplificando o fenómeno: uma pergunta *"Quem foi Karl Marx?"* exige como resposta uma passagem que contenha a expressão *"Karl Marx"*.[14] Para o padrão completo ser encontrado com sucesso na frase, este componente deve co-ocorrer com outro SN (que corresponde à resposta exacta), podendo surgir um a seguir ao outro (e.g. *"o filósofo Karl Marx"*) ou separados por uma vírgula ou verbo predicativo (e.g. *"Karl Marx, o intelectual alemão"* ou *"Karl Marx foi um revolucionário"*).

Para chegar a estas conclusões, foi feito um estudo das respostas a perguntas do tipo relevante. As perguntas e respectivas respostas estudadas foram aquelas que haviam sido compiladas para testar o sistema. Esta colecção foi construída com base nas perguntas utilizadas na

14 Note-se que, apesar de respostas contendo apenas *"Marx"* serem boas candidatas neste caso, o sistema não pode assumir que uma parte substitua o todo sem perda de especificidade.

conferência CLEF e será abordada em pormenor no próximo capítulo.

As respostas analisadas permitiram concluir que a quase totalidade dos casos era coberta por três tipos de construções, expostas nas seguintes linhas, onde **SN** representa um sintagma nominal e **V** um verbo predicativo (SN_1 e SN_2 correspondem aos sintagmas nominais da entidade-alvo e da resposta, podendo ocorrer invertidos, dado que a resposta pode surgir à esquerda ou à direita da entidade-alvo):

- SN_1 V SN_2
- SN_1, SN_2
- SN único contendo dois nomes: o alvo da pergunta e a respectiva resposta.

Reconhecimento de sintagmas nominais

Para reconhecer o SN que constitui a resposta curta, é usada outra ferramenta desenvolvida no grupo NLX, o LX-Tagger, que atribui categorias morfossintácticas às palavras que constituem uma frase [3]. Portanto, ao contrário do SN da pergunta, os SN que constituem respostas exactas são detectados através de métodos de PLN que apenas capturam expressões contendo palavras das classes morfossintácticas esperadas num SN.

Foi criada uma representação de sintagma nominal através de expressões regulares, usando as etiquetas do LX-Tagger que poderão fazer parte duma construção sintáctica desse tipo. O padrão que já se encontra em uso efectivo é o seguinte:

`QNT + ART|DEM + QNT + POSS + ORD + ADJ_PRE + CN|PNM + ADJ_POS`

QNT:	quantificador
ART:	artigo
DEM:	determinante demonstrativo
POSS:	determinante possessivo
ORD:	ordinal
ADJ_PRE:	adjectivo anterior ao nome
ADJ_POS:	adjectivo posterior ao nome
CN:	nome comum (possivelmente composto)
PNM:	nome próprio (possivelmente antecedido de CN ou título social)

Naturalmente, apenas o constituinte `CN|PNM` é obrigatório. Entretanto, está em curso a extensão e respectivos testes do padrão para permitir conjunções e sintagmas preposicionais.

A forma de utilização da ferramenta LX-Tagger a partir do Java é semelhante à descrita para o LX-Chunker na Secção 3.4.1, com a diferença de não ser necessária a criação de uma *thread* própria para passagem do texto à ferramenta (tanto a entrada como a saída da ferramenta realizam-se duma vez).

Capítulo 4

Avaliação

A pesquisa e adopção de medidas relativamente à avaliação do QueXting foram tarefas cuja importância se foi tornando evidente após os primeiros passos de desenvolvimento. Mais que uma simples validação do trabalho desenvolvido, a definição de um método de avaliação permitirá guiar o desenvolvimento futuro de forma sustentada, através da verificação dos resultados do sistema após alterações no seu funcionamento. Para atingir este objectivo, foram efectuadas pesquisas acerca de critérios, métodos e recursos de avaliação que se adequassem às características da aplicação. Foram também realizadas as primeiras avaliações de respostas curtas (num ambiente controlado) e de frases (na Internet).

4.1 Pontos-chave

4.1.1 Recursos

Para testar um SRP, é necessário um conjunto de perguntas e respostas. É também necessária uma colecção de documentos que contenham as respostas (não necessariamente a colecção usada originalmente pelo sistema).

Relativamente a conjuntos de perguntas e respostas, as alternativas são diversas:

- cartões de Trivial Pursuit;
- testes de escolha múltipla;
- conjuntos de FAQ *(frequently asked questions)* sobre determinado tema;
- páginas *web* contendo secções de perguntas ou votações efectuadas por utilizadores.

Estas fontes são restritas a determinado tema, mas também existe uma possibilidade interessante para sistemas de domínio aberto: os conjuntos de perguntas utilizados nas conferênci-

as de avaliação recentes. [12]

Colecções de teste reutilizáveis são difíceis de construir por duas razões. Primeiro, pelo facto de não existir uma forma canónica de resposta, não sendo possível determinar algoritmicamente se uma diferença entre duas sequências de caracteres é significativa. Segundo, porque os julgamentos devem ser feitos tendo em conta o documento particular de onde originou a sequência, o que pode levar a diferentes julgamentos para sequências idênticas.

As conferências de avaliação como a TREC e a CLEF disponibilizam os conjuntos de perguntas e respostas, e disponibilizam também o *corpus* de documentos onde as respostas devem ser procuradas. Para além destes, poderiam ser obtidos os padrões das respostas esperadas e listas de documentos relevantes para cada pergunta, para efectuar avaliações automáticas. Contudo, Lin [18] argumenta que estes últimos recursos ainda não se adequam à experimentação *post-hoc*, i.e. não produzem avaliações fiáveis de sistemas após as conferências. A justificação de Lin é a seguinte:

- As listas de documentos relevantes são obtidas recolhendo, para cada pergunta, o documento da primeira resposta de cada sistema. Com um documento por sistema, muitos documentos relevantes podem ficar de fora. Um sistema que venha a extrair uma resposta dum documento não listado pode não ser premiado adequadamente.
- Este facto é agravado pelos seguintes problemas:
  - Os métodos de identificação de documentos usados pelos sistemas são semelhantes, o que afecta a diversidade dos documentos que farão parte da lista.
  - Muitos sistemas deixam várias perguntas por responder, o significa que a lista não é aumentada nesses casos.

4.1.2 Critérios

Uma resposta deve ser dotada das características seguintes [12]:

- relevância (relativamente à questão);
- correcção (factual);
- concisão;

- completude;
- coerência (individual);
- justificação (pelo contexto do documento de origem).

A optimização de determinados critérios pode prejudicar outros, especialmente na avaliação automática. Neste processo, deve-se tentar garantir que os critérios se coadunam com o tipo de utilização e de utilizadores do sistema. Note-se que as opções por determinados critérios e meios de avaliação tem influência no desenvolvimento dos próprios sistemas, pois estes terão tendência a evoluir segundo os critérios definidos em vista de melhorarem a sua classificação em competições de anos sucessivos. Esta é, aliás, a desvantagem das conferências de avaliação: os sistemas participantes podem tornar-se demasiado afinados em determinados factores e menos noutros.

À medida que as próprias conferências evoluem, incorporam novos aspectos que influenciam a avaliação e consequentemente a evolução dos sistemas. Actualmente, os seus procedimentos já favorecem critérios não abrangidos inicialmente, tais como a concisão (através da limitação do tamanho das respostas) e a justificação (exigindo um suporte informativo adequado no documento fonte da resposta). Outros critérios, contudo, são mais difíceis de ter em conta – especialmente fora do ambiente controlado destas conferências – como são os casos da correcção factual e da completude.

4.1.3 Métricas

Finalizado o processo de avaliação individual das respostas, podem ser utilizadas métricas para classificar o sistema. Dado que nem sempre existe um conjunto finito e bem definido das respostas correctas que um sistema deveria encontrar, certas métricas clássicas da área de recuperação de informação, como a abrangência e a precisão, nem sempre são aplicáveis. Contudo, como veremos adiante, é possível aplicá-las ao nível das palavras que constituem as respostas, desde que exista o referido conjunto finito de respostas previstas.

A principal métrica usada nas conferências de avaliação é a fracção das questões correctamente respondidas (*accuracy*). Na aplicação desta métrica é normalmente avaliada apenas a primeira resposta do sistema a cada pergunta, pelo que existe interesse em complementá-la com outra métrica. Esta segunda métrica denomina-se *mean reciprocal rank* e avalia a posição

média (na lista de respostas de um sistema) em que é encontrada a primeira resposta correcta. Corresponde à média, sobre todas as questões, da fracção inversa da posição da resposta. Por exemplo, para uma determinada pergunta, se a primeira resposta correcta surgir na terceira posição da lista, o *reciprocal rank* é 1/3.

No caso de respostas longas, o princípio subjacente a esta métrica também pode ser aplicado ao nível das palavras que compõem as frases. Neste caso, a medida corresponde à fracção inversa da posição da primeira palavra na frase que faz parte da resposta propriamente dita. Esta métrica, em conjunto com a média do comprimento das respostas, permite medir o esforço do utilizador em discernir as respostas na frases apresentadas.

Para além da qualidade das respostas, um SRP que é disponibilizado *online* deve ser desenvolvido com especial atenção à usabilidade e à rapidez de resolução de pedidos. Deve ser tido em conta que um utilizador comum não desejará esperar muito mais que 10 a 20 segundos por uma resposta. Qualquer resposta que exija mais que meio minuto poderá ser inútil para utilizadores avançados da Internet. Actualmente, no Quexting, o tempo de execução do sistema para uma pergunta isolada – do tipo "Quem..?" (predicativa), para a qual é realizado processamento específico – encontra-se normalmente no intervalo de 7 a 15 segundos.

4.1.4 Avaliação automática

Nas primeiras conferências TREC que abordaram a tarefa de RP determinou-se que a consistência entre avaliadores (humanos) era suficientemente boa para possibilitar o emprego de apenas um avaliador, apesar de existirem pontualmente diferenças de opinião quanto à correcção de uma questão [31]. Este facto permite poupar recursos nas tarefas de avaliação, mas não resolve todos os problemas. Por exemplo, a presença do factor humano no processo impede os testes iterativos sistemáticos exigidos em técnicas de optimização automática (e.g. *hill-climbing,* algoritmos de aprendizagem automática) [12]. Mesmo sem utilizar este tipo de métodos, uma avaliação completamente automática tem uma nítida vantagem em custos temporais.

Infelizmente, as diferentes formas e componentes de respostas a uma pergunta arbitrária podem assumir enormes proporções. Certos sistemas chegam inclusivamente a produzir diferentes resultados em diferentes execuções. Estes factos dificultam bastante a avaliação e, em particular, a criação de métodos automáticos de avaliação. No caso mais restrito das perguntas

factuais, a avaliação automática torna-se mais tangível; de qualquer forma, não dispensa uma colecção completa (abrangente) de respostas correctas para cada pergunta do conjunto de testes. Esta tarefa fica facilitada se as respostas do sistema forem expressões curtas.

Para respostas curtas, é possível automatizar a comparação de respostas de um SRP relativamente a um conjunto de respostas criado por um especialista humano. Estas comparações já conseguem aproximar-se bastante (93–95%) daquelas efectuadas por humanos [12, 4]. Note-se que em experiências feitas na TREC a percentagem de acordo entre diferentes avaliadores humanos também foi 94% [31]... Contudo, a diferença da avaliação automática face à avaliação humana não se limita a uma ocasionalmente diferente aplicação de critérios, mas sim a erros sistemáticos com certos tipos de perguntas ou respostas, nomeadamente aquelas mais difíceis para o sistema.

Ainda assim, para efeitos de orientação interna de apoio ao desenvolvimento, a aplicação de métodos automáticos à avaliação de respostas curtas pode ser considerada suficientemente valiosa para servir de medida de evolução do sistema. A sua necessidade é inquestionável, dada a importância de saber se as alterações ao sistema durante o desenvolvimento produzem resultados e dado o tempo que exigem as avaliações manuais.

Trabalho relacionado

Uma referência na tarefa de avaliação automática de SRP é a ferramenta de avaliação automática Qaviar [4], que verifica a sobreposição das respostas obtidas com a resposta esperada, ao nível dos lemas das palavras. As respostas esperadas correspondem a conjuntos de palavras construídos manualmente com base em recursos da TREC e da Internet.

Outros sistemas aplicam técnicas de validação de respostas que exploram as relações semânticas entre resposta e pergunta, mas são de grande complexidade computacional e exigem muitos recursos linguísticos.

Uma abordagem diferente descrita em [19] corresponde a um método estatístico que usa as palavras-chave da pergunta em conjunto com as palavras-chave das respostas, com o intuito de validar uma resposta pela frequente co-ocorrência de ambas na Internet.

4.2 Avaliação do QueXting

4.2.1 Perguntas de teste

Como referido na Secção 4.1, uma fonte interessante de perguntas, respostas e documentos contendo estas respostas consiste nos recursos disponibilizados nas conferências de avaliação. O *corpus* de documentos-fonte em Português usado na CLEF consiste em conteúdo jornalístico proveniente de publicações do Público e da Folha de São Paulo de 1994 e 1995. Este *corpus* foi adoptado pelo QueXting para a criação de uma aplicação automática de testes.

À colecção de documentos destinada aos testes deve estar associada a uma colecção de perguntas cujas respostas estejam geralmente contidas nos documentos.[15] Com base nas perguntas usadas na CLEF desde a inclusão da língua portuguesa (2004-2006), foi criado um *corpus* de perguntas mantendo o formato XML usado em alguns dos recursos da conferência. A compilação detém toda a informação que existia associada a cada pergunta, com especial relevo para o tipo de resposta esperado e respostas existentes no *corpus* de documentos-fonte. As linhas do quadro seguinte exemplificam uma entrada típica do corpus de perguntas (com ênfase acrescentada à pergunta e às respostas previstas):

```
<q question_type="QUEM_PREDICATIVA" answer_type="PERSON"
category="D" cnt="216" restriction="NONE" source="CLEF"
year="2005">
<question>Quem foi Charles Bukowski? </question>
<answer docid="PUBLICO-940311-004" n="1">escritor</answer>
<answer docid="PUBLICO-940311-004" n="2">escritor norte-
americano de origem alemã</answer>
<answer docid="FOLHA-940311-107" n="3">escritor norte-
americano</answer>
</q>
```

Para além da informação originalmente associada a cada pergunta, certos campos foram acrescentados para uso no projecto QueXting, nomeadamente os campos `source`, `year` e `question_type` (este último referindo-se ao tipo de perguntas que o QueXting processa especificamente).

15 A possibilidade de perguntas sem resposta na colecção usada implica uma dificuldade extra para um SRP, pois este passa a precisar de medir o grau de confiança das respostas [12].

4.2.2 Automatização

Neste momento, o objectivo principal da tarefa de avaliação do QueXting é manter em funcionamento, durante o desenvolvimento, uma aplicação automatizada de testes. Esta aplicação permitirá, com custo temporal mínimo, observar a evolução do sistema após qualquer alteração (desde alterações mínimas de parâmetros do sistema até novas funcionalidades).

As respostas listadas na colecção de perguntas de teste, descrita nos parágrafos anteriores, são expressões exactas contendo apenas as palavras relevantes que um SRP deveria apresentar. A avaliação automática a aplicar ao QueXting focar-se-á, portanto, nas respostas curtas, para as quais será possível medir a abrangência e precisão ao nível das palavras contidas na resposta. Estas medidas são aplicáveis por se conhecerem à partida as respostas que existem nos documentos-fonte e as palavras que as constituem.

Naturalmente, as perguntas da colecção poderão ser usadas também para avaliação sobre a Internet, no entanto as respostas possíveis deixam de estar limitadas àquelas listadas para cada pergunta pelos organizadores da CLEF. Dada a ausência de um conjunto finito de padrões de respostas e dada a incerteza sobre a existência e permanência de documentos com os conteúdos pretendidos, não é possível realizar uma avaliação na rede usando métodos automáticos. Portanto, a abordagem automática usará apenas os documentos da CLEF.

Aparte esta abordagem automática, é necessário manter mecanismos de verificação do funcionamento do sistema na Internet, dada a diferença entre os documentos da CLEF e aqueles que se encontram na rede. A complexidade e estrutura variável destes últimos são factores que podem ter influência no funcionamento do sistema, pelo que eventualmente podem existir problemas que se manifestem exclusivamente na execução *online*.

Casos problemáticos

Apesar de evitar a repetição custosa de análises manuais, a avaliação automática levanta alguns problemas específicos, que deverão ser tidos em conta. Breck et al. [4] referem diversos casos que exigem especial atenção, dada a rigidez da avaliação automática. Os pontos seguintes constituem exemplos importantes para cada um desses casos:

- Expressões numéricas e temporais:

- a expressão *10%* deve ser comparável a *10 porcento* ou *dez porcento;*
- datas relativas (e.g. *há 20 anos atrás*) devem ser resolvidas antes de comparadas com datas absolutas.

- Nomes parciais:
  - frequentemente, os nomes compostos são formados por partes de importância variável (p.e. um apelido pode ser a parte mais importante do nome próprio de uma pessoa) – este facto implica que certas respostas deveriam ser consideradas correctas, apesar de parciais.
- Granularidade:
  - o grau de especificidade que torna uma resposta útil é variável.
- Contexto:
  - falta de contexto adequado (geográfico, temporal, etc.) na pergunta.
- Palavras descartadas:
  - certas expressões podem coincidir com palavras descartadas pelo sistema.
- Forma lógica (semântica):
  - duas frases com as mesmas palavras não são necessariamente iguais (*o homem mordeu o cão* não tem o mesmo significado que *o cão mordeu o homem*);
  - a negação lógica não é tida em conta pelas métricas automáticas.
- Questões mais complexas:
  - quanto mais aberto for o âmbito de uma questão, mais possibilidades de resposta poderão existir, não sendo garantido que uma chave de resposta (p.e. implementada sob a forma de lista de palavras ou tópicos) seja totalmente abrangente.

Alguns destes problemas podem ser resolvidos especificando os padrões de respostas de forma a diferenciar palavras mais importantes, permitir gamas numéricas em lugar de valores exactos ou incluir restrições sobre a resposta (p.e. restrições temporais).

4.2.3 Primeira avaliação de respostas curtas

Os padrões desenvolvidos para detectar respostas curtas às perguntas *"Quem..?"* predicativas já capturam algumas expressões úteis contidas nas frases devolvidas pelo QueXting. Dados os recursos de avaliação compilados, nomeadamente as perguntas da CLEF com respostas previstas associadas, é possível efectuar uma avaliação automática ao nível da abrangência e precisão das respostas curtas obtidas pelo QueXting.[16]

Metodologia

Foram adoptadas as 68 perguntas do tipo *"Quem..?"* (predicativas) existentes na colecção de perguntas de teste que tinham respostas previstas no *corpus* de documentos da CLEF. Dada a relativa escassez de perguntas com resposta associada, todas as perguntas disponíveis foram usadas para testes, não existindo uma separação entre conjunto de teste e conjunto de desenvolvimento. Uma listagem das perguntas testadas encontra-se disponível no Anexo D.

Para cada questão, analisaram-se as 5 primeiras frases obtidas como resposta: as respostas curtas detectadas nestas frases foram automaticamente comparadas, palavra a palavra, com as respostas previstas, com o objectivo de medir abrangência e precisão. No contexto desta avaliação, para cada resposta curta, as fórmulas usadas para calcular estas medidas foram as seguintes:

- $\text{abrangência} = \dfrac{\textit{número de palavras relevantes obtidas}}{\textit{número de palavras relevantes previstas}}$
- $\text{precisão} = \dfrac{\textit{número de palavras relevantes obtidas}}{\textit{número de palavras obtidas no total}}$

Foi também calculada a denominada medida-F, que combina abrangência e precisão num só valor. Para a aplicação desta métrica neste contexto de avaliação, tanto a abrangência como a precisão são igualmente importantes: a primeira para completude da resposta; a segunda para concisão (não interessa obter expressões demasiado longas pois já se apresenta a frase completa após a expressão curta). Deste modo, a fórmula usada para cálculo da medida-F foi a tradicional, que dá um peso igual aos componentes de abrangência e precisão, e resulta na se-

16 Um método automático de avaliação de respostas curtas foi implementado por Pedro Martins, do grupo NLX. A sua preciosa ajuda estendeu-se à execução deste método sobre as perguntas de teste, à avaliação de frases na Internet (descrita na secção seguinte) e à compilação dos resultados destas duas avaliações.

guinte expressão:

- $F = \frac{2 * abrangência * precisão}{abrangência + precisão}$

Dado que a mesma pergunta pode ter várias respostas certas previstas e que diversas frases-resposta para a mesma pergunta podem conter expressões curtas adequadas, para efeitos de média final dos valores destas métricas contabilizou-se apenas uma resposta curta por questão: aquela com o melhor valor de medida-F relativamente a alguma das respostas listadas como certas.

A posição em que ocorre cada uma das respostas contabilizadas para os valores médios foi usada para calcular o *mean reciprocal rank*. Através destes procedimentos, é possível medir se o sistema está a capturar as melhores respostas e também se a posição delas é adequada (idealmente, em primeiro lugar na lista).

Resultados

Foram obtidas respostas curtas em 32 das 68 perguntas pelo que, à partida, um máximo de 47% das perguntas lançadas resultaram numa resposta curta adequada. Para conseguir um maior número de respostas curtas, terá de ser efectuado um estudo das frases onde a detecção de padrões sintácticos não foi bem sucedida e, se necessário, implementar a detecção de novos tipos de padrões. Para medir até que ponto as respostas obtidas foram realmente adequadas, aplicaram-se as métricas referidas acima. Os valores médios obtidos destas medições são apresentados na Tabela 3:

	Abrangência	Precisão	Medida-F
Valor médio	73,41%	92,69%	78,52%
Desvio-padrão	33,29%	24,65%	29,02%

Tabela 3: Valores médios de abrangência, precisão e medida-F das respostas curtas obtidas

Os valores obtidos demonstram que as melhores respostas curtas obtidas já são interessantes. A precisão é elevada, dado que os padrões usados para capturar expressões curtas ainda não detectam sintagmas nominais complexos que envolvam conjunções ou sintagmas preposicionais. Desta forma, as expressões obtidas são realmente curtas na grande maioria dos casos. Apesar da concisão das respostas, o valor de abrangência também não é desprezável.

Infelizmente, nos casos em que faltam palavras à resposta, estas palavras podem ser essenciais para uma correcta interpretação. Consideremos um exemplo deste fenómeno:

- Pergunta: *"Quem é Tenzin Gyatso?"*
    - Frase 1: "*Tenzin Gyatso é o líder espiritual do Tibete*"
    - Frase 2: *"O líder espiritual do Tibete, Tenzin Gyatso (...)"*

Na frase 1, a resposta curta detectada na frase seria *líder espiritual*, o que consistiria numa resposta correcta, embora incompleta. Contudo, na frase 2, a expressão curta extraída seria apenas *Tibete*, o que constitui uma resposta efectivamente errada. Cada uma das expressões é capturada por constituir um sintagma nominal válido (embora englobado noutro de maior dimensão, por meio de uma preposição). A falta de suporte a sintagmas preposicionais revela-se particularmente problemática em casos como os da frase 2, em que a resposta curta ocorre à esquerda da entidade-alvo.

A Tabela 4 seguinte representa a posição das melhores respostas curtas por pergunta. O *mean reciprocal rank* obtido a partir dos valores apresentados é 0,91. É natural que a maioria das respostas curtas ocorram nas primeiras frases da lista, pois o QueXting valoriza as frases onde foram encontradas respostas curtas.

Posição	***Respostas curtas***
1	27
2	3
3	1
4	0
5	1

Tabela 4: Posição das melhores respostas curtas por pergunta

Uma das respostas curtas analisadas é de particular interesse. A resposta listada como correcta é *cosmonauta russo* mas, nos documentos-fonte, ocorre apenas a expressão *cosmonautas russos*. Era esperada do sistema a capacidade de abstrair ou converter o número deste sintagma nominal. Para a resposta obtida pelo QueXting (*cosmonautas russos*) a abrangência e a precisão foram contabilizadas como zero.

Podem ser tomadas acções para resolver este tipo de problemas. Na ferramenta de avaliação automática Qaviar, descrita anteriormente, as medidas de abrangência e precisão não são calculadas directamente sobre as palavras, mas sobre os lemas destas. Por exemplo, a resposta *"Fisherman: They called it El Niño"* é convertida {fisherman call niño}. [4]

Com esta abordagem de avaliação, o caso referido deixaria de ser problema pois a resposta seria convertida em *cosmonauta russo.* Contudo, o ideal não é realizar uma avaliação permissiva (que pode ter as suas vantagens), mas sim implementar um sistema robusto que saiba detectar os moldes em que deve devolver uma resposta ao utilizador.

Concluindo, o método de avaliação de respostas curtas aqui descrito, devido à automatização do processo, permite aplicar sem esforço métricas conhecidas. Avaliações futuras poderão ser comparadas a esta em termos de:

(a) percentagem de perguntas às quais se obtêm respostas curtas;

(b) abrangência e precisão destas respostas curtas.

Por outro lado, a avaliação manual na Internet poderá cobrir aspectos complicados, nomeadamente o abordado nos parágrafos anteriores e outros casos problemáticos referidos na Secção 4.1.

4.2.4 Primeira avaliação na Internet

Por estas razões e para ter uma noção concreta da capacidade actual de obtenção de respostas a partir da rede, realizou-se uma primeira avaliação do sistema no seu ambiente natural, abordando o tipo de pergunta para o qual já existe suporte específico.

Deve ser tido em conta que não foi possível efectuar qualquer estudo prévio sobre a existência de respostas na Internet para as perguntas utilizadas, que foram criadas com base numa colecção fechada de documentos de origem jornalística. Este facto implica que esta avaliação terá fortes limitações em termos das conclusões que poderão retirar. Apenas se for feito um estudo sobre a existência de respostas na rede para as perguntas utilizadas é que os resultados obtidos se poderão considerar bons ou maus relativamente às perguntas lançadas. Pela natureza dinâmica da rede, sucessivas avaliações deste tipo também não poderão ser comparadas com o mesmo nível de confiança que se terá na avaliação de respostas curtas, que é efectuada

num ambiente controlado.

Metodologia

O processo de avaliação na Internet envolveu as seguintes tarefas:

(1) adopção das 68 perguntas do tipo *"Quem..?"* (predicativas) que também foram utilizadas na avaliação de respostas curtas que foi executada sobre os documentos da CLEF;

(2) lançamento das perguntas ao sistema, com 30 a 60 segundos de intervalo, e salvaguarda das respostas;

(3) análise manual das 2 primeiras frases obtidas como resposta.

Para cada uma das 68 perguntas lançadas, foram analisadas e contabilizadas todas as respostas existentes nas 2 primeiras posições. A tarefa de análise de respostas consistiu na classificação de cada resposta como *correcta, inexacta, incompleta* ou *errada.* A categoria *inexacta* cobre os casos em que a resposta é completa mas surge num contexto errado ou perdida numa frase demasiado comprida.

Consideremos um exemplo real destes julgamentos, dada a pergunta "*Quem foi o sucessor de Kim Il Sung?"*:

- Resposta correcta:
    - *Devido às suas grandes contribuições,* ***Kim Jong Il*** *foi o sucessor de Kim Il Sung na liderança do país e, em 1997, foi eleito secretário geral do Partido do Trabalho da Coréia.*
- Resposta inexacta:
    - *Em uma frase, poderia se dizer que a contribuição de Kim Il Sung à Humanidade foi sublinhar, como ninguém antes dele, o papel decisivo e imprescindível da consciência e da unidade na luta pela libertação do jugo imperialista, contribuição que ele deu com seus escritos teóricos, desenvolvidos depois por* ***Kim Zong Il*** *- seu filho e sucessor -, e com sua vida, de armas na mão contra os ocupantes da sua Pátria, como governante, dirigente do Partido do Trabalho da Coréia e grande líder da Revolução Coreana.*

- Resposta incompleta (por não indicar que a sucessão foi efectiva):
    - *Sabe-se que **Kim Jong-il** passou décadas a ser preparado para suceder a Kim Il-sung.*
- Resposta errada:
    - *No tempo de Kim il-Sung, que iniciou a Guerra da Coreia (a que se refere esta pintura de propaganda), Moscovo e Pequim pareciam ter mais influência.*

As respostas não foram sujeitas a validação temporal pelo que diferentes respostas foram consideradas correctas, desde que justificadas pelo contexto. Esta verificação estaria para além do âmbito do trabalho abordado nesta dissertação. Este facto favorece a contagem de respostas correctas; contudo outro critério tem o efeito contrário. Foram detectadas algumas respostas noutras línguas (Castelhano e Inglês) e, quando correctas, foram consideradas inexactas por exigirem conhecimento do idioma para sua interpretação.

Resultados

As respostas analisadas resultaram de um total de 68 perguntas do tipo "Quem..?" (predicativas). Apenas uma das 68 perguntas ficou sem qualquer resposta; para outras duas perguntas, apenas foi encontrada uma resposta. A Tabela 5 seguinte representa a contagem de respostas correctas, inexactas, incompletas e erradas, encontradas nas duas primeiras posições da lista de respostas devolvida pelo QueXting:

Posição	Correctas	Inexactas	Incompletas	Erradas	TOTAL
1	12	16	4	35	67
2	10	14	10	31	65
TOTAL	22	30	14	66	132

Tabela 5: Contagem de respostas correctas, inexactas, incompletas e erradas

O total de respostas analisadas foi 132: destas, 22 foram consideradas correctas, 30 inexactas, 14 incompletas e 66 erradas. Foram obtidas poucas respostas absolutamente correctas, embora haja um número significativo de respostas inexactas que poderão conter expressões curtas adequadas. O gráfico da Figura 10 representa a distribuição de respostas por posição:

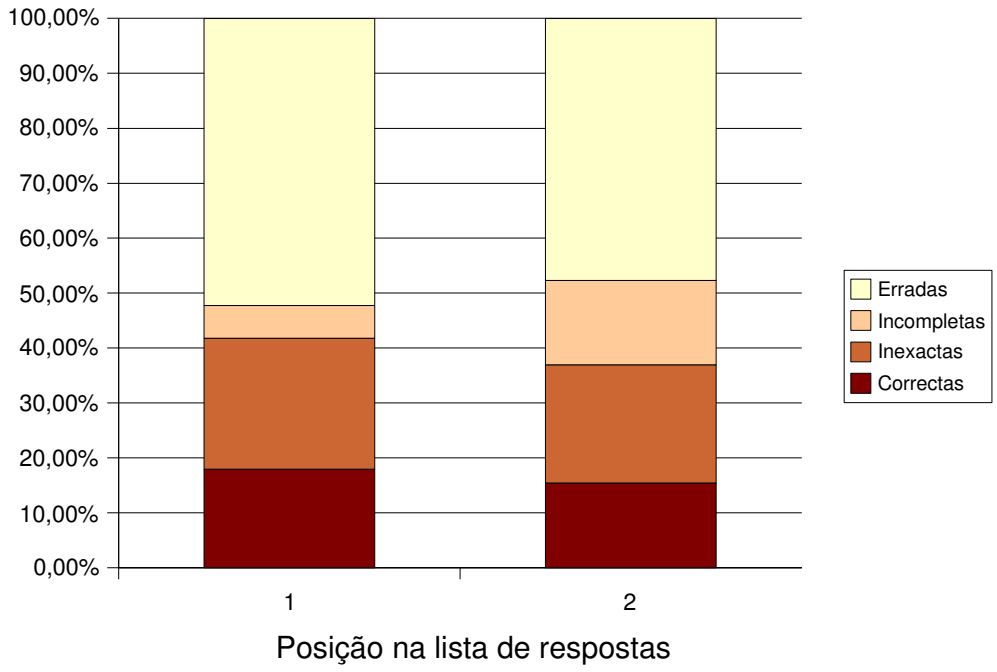

Figura 10: Distribuição de respostas por posição

Na primeira posição, cerca de 18% das respostas obtidas foram correctas. No total das 2 posições obtém-se um número significativo de respostas absolutamente correctas (16-18%) e cerca de 30% de respostas parciais (inexactas ou incompletas). Portanto, aproximadamente 50% das respostas obtidas nas duas primeiras posições têm algum conteúdo útil.

Quanto à percentagem de respostas erradas (os restantes 50%), poderá ser reduzida com a introdução de processamento semântico. O reconhecedor de entidades nomeadas, associado ao dicionário de perguntas/respostas, permitirá filtrar respostas de tipos inadequados. Além disso irá favorecer as respostas do tipo adequado, o que deverá trazer mais respostas correctas para estes primeiros lugares do *ranking*.

A distribuição de respostas correctas pela posição – em maior número na primeira do que na segunda posição – sugerem que o sistema está a conseguir valorizar e posicionar as respostas correctas de forma adequada. Este fenómeno verifica-se também ligeiramente com as respostas inexactas (que poderiam ser correctas em contextos adequados). Note-se que nas respostas inexactas poderão existir expressões curtas que constituam uma boa resposta. Este factor será mais valorizado na avaliação automática de respostas curtas.

Os resultados obtidos não podem ser comparados directamente com os de outros sistemas

pois (i) a avaliação foi efectuada isoladamente, (ii) apenas sobre um tipo de pergunta, (iii) com um conjunto de perguntas único e (iv) sem verificação da existência de respostas. Contudo, se houvesse resposta garantida às perguntas lançadas, o objectivo natural seria atingir ou exceder a marca dos 70% de respostas correctas na primeira posição da lista devolvida pelo sistema, algo que é conseguido pelos melhores sistemas, em execuções controladas, para todas os vários tipos de perguntas factuais da TREC ou da CLEF.

4.2.5 Avaliação comparativa

As avaliações realizadas em conferências foram, desde as primeiras edições, alvo de um processo próprio de validação da metodologia utilizada. Este processo confirmou a estabilidade dos resultados e a validade do processo. Contudo, como referido na Secção 4.1 sob a entrada *Recursos*, as comparações *a posteriori* com os sistemas que concorreram ao evento não são válidas. Portanto, a melhor forma de colocar o QueXting à prova face a outros SRP será através de participações na CLEF. Esta opção permitirá fazer parte de um processo normalizado, em concorrência com vários sistemas em simultâneo.

Um processo alternativo é a realização de experiências comparativas na Internet. Aqui, o sistema está no seu ambiente natural de execução. Contudo, dados os custos temporais de preparação e execução desta tarefa e a dificuldade na definição e verificação de quais podem ser as respostas correctas para cada pergunta, este método terá de ser estudado adequadamente e utilizado apenas em circunstâncias de alguma estabilidade no desenvolvimento.

Poderiam ser realizadas comparações face ao AnswerBus, que suporta perguntas em Português, mas este sistema responde em Inglês e a partir da informação de documentos nessa língua. Este último factor implica mesmo que muitas questões não deverão ter qualquer resposta. O Esfinge é o único SRP alternativo adequado ao Português e disponível na rede. Este sistema poderá fazer parte de uma avaliação manual comparativa assim que o nível e a estabilidade de desenvolvimento do QueXting o justifiquem.

Capítulo 5

Conclusão

5.1 Principais resultados

O trabalho apresentado nesta dissertação teve como resultado principal a infra-estrutura básica do SRP QueXting, assente em 3 módulos: processamento de questões, recolha de documentos relevantes e extracção de respostas. Como exercício experimental, para verificação do princípio ao fim do processo de funcionamento do sistema, foi implementado o componente específico para as perguntas do tipo *"Quem..?"* predicativas. A esta infra-estrutura acrescem as páginas de descrição do projecto e de uma interface para teste do sistema. Por último, recursos e resultados preliminares de avaliação foram compilados.

5.1.1 Infra-estrutura

O trabalho realizado ao nível da infra-estrutura básica divide-se pelos 3 módulos da arquitectura descrita no Capítulo 3.

Processamento de questões

A principal funcionalidade geral (independente do pedido) implementada foi:

- Simplificação do pedido do utilizador com vista à manutenção das palavras-chave que devem ser fornecidas aos módulos de recolha de documentos e de extracção de respostas.

Recolha de documentos-fonte

- A identificação de documentos relevantes para cada pergunta é efectuada através de motores de busca de documentos existentes *online*. Para tal, são lançados pedidos aos

motores contendo palavras-chave da pergunta do utilizador e são recolhidos os endereços dos documentos listados como resultado pelos motores. Estes documentos são descarregados dos respectivos servidores na sua totalidade para processamento directo.

- Uma das principais funcionalidades adicionadas a este módulo consistiu na interpretação correcta dos caracteres dos documentos. Cada página da Internet pode estar escrita com um esquema diferente de codificação de caracteres (*character encoding*). Portanto, para interpretar e representar correctamente caracteres acentuados e simbólicos, tornou-se necessário reconhecer o esquema de codificação de cada página descarregada.
- A análise do código HTML que constitui as páginas foi também alvo de alguns melhoramentos, visando evitar erros do analisador utilizado (o da biblioteca *swing* do Java).
- Foi ainda implementado um método simples de detecção de documentos duplicados alojados em URL's diferentes, baseado em excertos (dos documentos) oferecidos pelos motores de busca nas suas páginas de resultados.

Extracção de respostas

- Segmentação do texto dos documentos-fonte em frases, usando a ferramenta LX-Chunker do NLX. O facto de esta ferramenta esperar texto bem formado obriga a lidar previamente com todas as quebras de linha existentes no código HTML dos documentos. Esta medida evita a colagem de frases sem pontuação final adequada.
- Valorização preliminar de frases, baseada no número e na co-ocorrência das palavras-chave encontradas.

5.1.2 Processamento de perguntas "Quem..?" predicativas

A necessidade de adicionar capacidades de processamento de tipos particulares de perguntas implicou o desenvolvimento de funcionalidades adicionais nos módulos de processamento de questões e extracção de respostas.

Processamento de questões

- Identificação do tipo de pergunta a partir do seu padrão sintáctico: implementada para as perguntas "Quem..?" predicativas. Os factores essenciais para a detecção deste tipo de pergunta correspondem à presença do pronome interrogativo *quem* e de um verbo predicativo.
- Salvaguarda do sintagma nominal que se segue ao verbo predicativo, que fará parte dos padrões sintácticos esperados nas respostas.

Extracção de respostas

- Criação e detecção de padrões sintácticos esperados nas respostas ao tipo de pergunta referido. As passagens das frases encontradas através destes padrões constituem respostas curtas que são apresentadas ao utilizador antes da respectiva frase de origem.
- A detecção de padrões adequados às perguntas do tipo *"Quem..?"* envolveu o reconhecimento de sintagmas nominais. Para tal, as frases são processadas pelo LX-Tagger, uma ferramenta que associa a cada palavra a sua categoria morfossintáctica. Foi definido um padrão de sintagma nominal composto pelas categorias morfossintácticas relevantes.

5.1.3 Páginas web

Foram criadas duas páginas *web* criadas no âmbito do QueXting. Uma página disponibiliza na rede um resumo sobre o projecto QueXting. O logótipo criado para esta página poderá ser adoptado também para a interface de utilizador. A página de interface consiste numa versão interna de testes, permitindo introduzir a pergunta, seleccionar os motores de busca de documentos e visualizar informação relativa ao funcionamento dos 3 módulos do sistema.

5.1.4 Notas, recursos e resultados de avaliação

Dada a natureza complexa da aplicação em desenvolvimento e a dificuldade em ter uma noção abrangente acerca dos avanços produzidos por sucessivas alterações, tornou-se premente definir estratégias e meios de avaliação. Após alguma investigação, foram compiladas algu-

mas notas sobre recursos, critérios e métricas aplicáveis ao QueXting.

Os principais recursos utilizados são um conjunto de perguntas compilado a partir dos recursos da CLEF, assim como o *corpus* de documentos-fonte utilizado nessa conferência. Estes recursos permitem uma avaliação de respostas curtas que, no futuro, poderá ser realizada de forma automática. Tomou-se nota de alguns casos potencialmente problemáticos na automatização da avaliação, que deverão ser tidos em conta. Finalmente, foram realizadas duas avaliações preliminares do sistema: uma incidindo sobre as respostas curtas, utilizando apenas os documentos da CLEF como fonte; outra ao nível de frases completas, utilizando apenas a Internet como fonte.

As perguntas testadas foram do tipo "Quem..?" predicativas, para as quais existe já suporte específico. Consideraram-se as 68 perguntas desse tipo para as quais se sabia existirem respostas nos documentos da CLEF (embora sem ter o mesmo conhecimento no que diz respeito à Internet). Foram conseguidas respostas curtas em 32 das 68 perguntas testadas sobre os documentos da CLEF, tendo as melhores respostas curtas por pergunta atingido um valor médio de medida-F de 78,52%. Para a avaliação na Internet não é possível realizar um estudo das perguntas que teriam efectivamente resposta disponível. Ainda assim foram testadas as mesmas 68 perguntas e analisadas as duas primeiras respostas a cada uma. Obteve-se, nas duas primeiras posições da lista devolvida, 16-18% de respostas correctas e cerca de 30% de respostas parciais.

Os *rankings* das respostas correctas obtidas, tanto na primeira experiência como na segunda, sugerem que o sistema está a valorizar adequadamente as respostas correctas que encontra. Este fenómeno deverá ainda melhorar com a utilização do reconhecedor de entidades nomeadas, que irá filtrar as respostas de tipos inadequados. Quanto à experiência de avaliação na *web*, os resultados sugerem que este tipo de avaliação só deverá ser feita em situações de estabilidade no desenvolvimento (também porque uma experiência adequada exige uma verificação da existência de respostas para as perguntas testadas). Em termos de avaliações comparativas que envolvam outros SRP, a melhor solução será a participação nas conferências de avaliação.

5.2 Trabalho futuro

Neste tipo de aplicação existirão sempre muitas possibilidades/necessidades de desenvolvimento adicional. No estado actual, destacam-se aquelas descritas nos pontos seguintes.

- Suporte a mais tipos de perguntas (que deverá começar pelas *"Quando..?"*).
- Extensão do padrão usado para detecção de sintagmas nominais, de forma a incluir conjunções e sintagmas preposicionais (permitindo encontrar expressões curtas como *"líder político e espiritual do Tibete"* em lugar de simplesmente *"líder político"*).
- Verificação de concordâncias de tempos verbais e traços nominais de género e número (permitindo descartar respostas com características díspares relativamente à pergunta).
- Verificação de outras restrições temporais da pergunta e respectivas respostas (nomeadamente quando a pergunta se refere a determinado período temporal, ou a resposta contém alguma informação temporal que deveria ser tida em conta).
- Uso do reconhecedor de entidades nomeadas em desenvolvimento no NLX. Uma taxonomia própria de tipos de respostas poderá ser construída e aplicada num dicionário de perguntas/respostas.
    - Esta medida reforçaria a detecção e confiança em respostas curtas contendo entidades ou eventos de tipo adequado; por exemplo respostas consistindo numa *pessoa* ou *organização* a perguntas do tipo *"Quem..?"*.
    - Estas ferramentas poderão ser complementadas com ontologias de semântica lexical que permitiriam raciocínios através das relações entre palavras.
- Processamento adicional das perguntas:
    - Expansão dos pedidos aos motores de busca com palavras sintáctica ou semanticamente relacionadas; lematização com vista à utilização de formas canónicas das palavras.
    - Manutenção de todas as palavras encontradas numa entidade nomeada.
- *Parser* HTML:
    - O analisador utilizado (existente na biblioteca *swing* do Java) revelou algumas li-

mitações, nomeadamente na detecção de código de *script* ou estilo CSS, e na detecção de etiquetas de fecho do tipo />. Existem *parsers* alternativos que poderão ser explorados.[17]

- Cruzamento das respostas curtas obtidas de modo a *fundir* aquelas que são iguais e aumentar o seu valor, listando após a resposta curta todas as frases que a contêm.
- Disponibilização do serviço *online* para uso generalizado, assim que suporte mais tipos de perguntas e use o reconhecedor de entidades nomeadas (dado que estas deverão ser as alterações com maior impacto nos resultados).
- Considerar o uso de diferentes motores de busca de documentos, que permitam encontrar notícias recentes ou preços e características de produtos de consumo.
- Detecção de documentos idênticos:
  - A detecção de documentos idênticos alojados em URL's diferentes baseia-se em excertos dos documentos fornecidos pelos motores de busca. Este método é falível, pelo que se deve testar o desempenho de métodos alternativos (p.e. efectuando uma comparação das palavras menos frequentes de cada documento e definindo uma percentagem mínima de semelhança).
- *Cache* de perguntas frequentes, melhorando o desempenho nestes casos e aliviando a carga do sistema em momentos de carga elevada.
- Participação em conferências de avaliação como forma de validação e comparação com outros sistemas.

5.3 Comentário crítico

O estado actual de maturidade do sistema não justifica a sua disponibilização *online*, mas a continuação do projecto com suporte a novos tipos de perguntas e a integração do reconhecedor de entidades nomeadas deverão alterar esta realidade.

O trabalho desenvolvido ao nível da infra-estrutura do QueXting e a investigação realizada sobre as possibilidades de avaliação constituem bases fundamentais e valiosas para uma rápida evolução e teste das capacidades do sistema a partir deste ponto. Existem alguns factores a

17 Uma fonte possível é o endereço: http://www.java-source.net/open-source/html-parsers

desenvolver e testar na própria infra-estrutura; contudo, dada a complexidade da aplicação, tornou-se fundamental criar uma plataforma de base que permitisse a execução do sistema do princípio ao fim e o posterior incremento das suas capacidades.

Após a concretização desse objectivo procedeu-se ao enriquecimento do suporte a perguntas específicas. Este passo foi concretizado apenas para um tipo de pergunta, mas espera-se que a repetição do processo para diferentes tipos seja mais célere, pois parte do trabalho será reaproveitado.

Não puderam ser extraídas conclusões definitivas sobre a capacidade de obtenção de respostas a partir da *web* para o tipo de pergunta testado, mas a capacidade de detecção de expressões curtas de resposta é promissora (cf. Figura 11) apesar de ainda estar longe de uma cobertura total. A aplicação de ferramentas de processamento semântico (o reconhecedor de entidades nomeadas e recursos associados) poderá confirmar esta expectativa.

```
----------------------
RESPOSTA #1 (valor 6)
----------------------

Resposta exacta: ator perfeito
Resposta exacta: cristalina
Frase completa: Verdade pura e cristalina, Edward Norton é um ator perfeito e não está para brincadeira.

Data: Mon Nov 15 19:55:37 WET 1999
Documento: www.geocities.com/Hollywood/Theater/3451/spotlight.html
Motor: MSN (rank 1)

Busca e parse de resultado: 1011 ms.
Download e parse de documento: 952 ms.
Extraccao de resposta: 10 ms.
Tempo total: 2373 ms.
```

Figura 11: Primeiras respostas exactas do QueXting à pergunta: "Quem é Edward Norton?"

A criação de um dicionário de perguntas/respostas deverá ser despoletada pela aplicação do reconhecedor de entidades nomeadas e de ontologias de semântica lexical. A aplicação destas ferramentas (em desenvolvimento no NLX) é um processo complexo e depende do seu estado de maturidade, pelo que ainda não foi possível realizá-la. Foi sim realizada uma análise dos tipos de perguntas factuais possíveis, que neste momento guia a escolha dos tipos para os quais se pretende implementar suporte específico.

A necessidade de definir recursos e métodos de avaliação levou a duas medidas adicionais: uma pesquisa sobre metodologia adequada à avaliação do SRP QueXting e a realização de duas experiências de avaliação que produziram informação útil sobre os resultados obtidos

pelo sistema e sobre a própria metodologia utilizada.

Em termos de desempenho temporal, o tempo de execução do QueXting para uma pergunta isolada do tipo "Quem..?" (predicativa) é de 7 a 15 segundos. Pode ser considerado um valor positivo tendo em conta que inclui o descarregamento de documentos da Internet. Contudo, o uso das ferramentas adicionais referidas anteriormente deverá aumentar este valor. Assim, para além da avaliação dos resultados obtidos com essas ferramentas, tornar-se-á essencial verificar o impacto temporal. Qualquer resposta que exija mais que meio minuto poderá ser inútil para utilizadores avançados da Internet. Se este limite for ultrapassado, pode ser considerada a hipótese de usar menos motores de busca de documentos por pedido, incluindo algum método dinâmico de escolha do motor consoante a pergunta efectuada ou as palavras nela contidas.

Para finalizar, um comentário sobre o uso da Internet para RP. O tamanho da rede e a variedade de documentos são factores que permitem obter respostas a perguntas muito variadas. A redundância de informação é uma vantagem quando consiste em diferentes abordagens ou contextos para um mesmo tema. Por outro lado, o SRP está dependente da capacidade de identificação de documentos-fonte (que, no caso do QueXting, se baseia em motores de busca já existentes na rede). Isto significa ter de lidar com as restrições destes motores ao nível de pedidos, sintaxe, consistência de resultados, transparência, etc. [15]. A respeito desta dependência de recursos das grandes companhias, reaproveito as palavras expectantes de Kilgariff: *também nós* [no meio académico] *precisamos de criar recursos a esta escala e disponibilizá-los aos investigadores* [15], mesmo que isso implique esforços adicionais de cooperação! Para tal, há que tornar os recursos científicos competitivos face aos recursos dessas companhias, e disponibilizá-los ao restante meio científico.

ANEXO A

Lista de serviços online de resposta-a-perguntas

Activos	Autores/organizações	Endereços
AnswerBus	Zhiping Zheng	http://answerbus.coli.uni-saarland.de/
Arizona State Univ.	Dmitri Roussinov	http://qa.wpcarey.asu.edu/
Ask	IAC Search & Media	http://www.ask.com/
askEd!	Ed Whittaker	http://asked.jp/
Brainboost	Answers Corp.	http://www.brainboost.com/
DKFI	DKFI	http://experimental-quetal.dfki.de/redirect.jsp
Esfinge	Linguateca	http://www.linguateca.pt/Esfinge/
LCC	Language Computer Corp.	http://www.languagecomputer.com/
NSIR	CLAIR (Michigan Univ.)	http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi
START	MIT/AI Lab	http://start.csail.mit.edu/
TellMe	Luiz Pizzato (Macquarie Univ.)	http://www.ics.mq.edu.au/~pizzato/tellme

Indisponíveis	Autores/organizações	Endereços
IONAUT	Steve Abney, Michael Collins	http://www.ionaut.com:8400/
LAMP	Universidade de Singapura	http://hal.comp.nus.edu.sg/cgi-bin/smadellz/lamp_query.pl
Powerset	Powerset	http://www.powerset.com/
QuASM	UMASS/CIIR	http://ciir.cs.umass.edu/~reu2/
SiteQA Demo	DiQuest.com Inc.	http://ressell.postech.ac.kr/~pinesnow/siteqeng/

Motores de busca de documentos mencionados na dissertação:

Live (MSN)	http://www.live.com
Google	http://www.google.com
Yahoo!	http://www.yahoo.com

ANEXO B

Text Retrieval Conference: Question Answering track

A TREC é realizada anualmente nos E.U.A desde 1992, com o propósito principal de desenvolver a pesquisa sobre recuperação de informação, a sua avaliação e aplicação prática [12]. A conferência envolve uma série de *workshops* com o objectivo de realizar testes de larga escala e trocas de ideias sobre a tecnologia de recuperação de informação, sendo a tecnologia de RP alvo de um destes *workshops* [34].

O foco e a evolução deste evento em anos recentes permitem discernir o estado da arte. Em termos de resultados, os melhores valores têm-se aproximado dos 70% das perguntas factuais correctamente respondidas, embora sem grande evolução ao longo dos últimos anos, pois a tecnologia começa a amadurecer e os pequenos melhoramentos de ano para ano são compensados pelo endurecer de restrições relativamente ao formato das respostas ou pelo teste de perguntas ligeiramente mais complexas.

2003

No ano de 2003, o *workshop* TREC/QA envolvia 2 tarefas distintas [32]:

- ***Passages***: Dedicada a perguntas factuais. A resposta esperada era um extracto de até 250 caracteres que devia tornar clara a resposta, sem ambiguidades (e.g. *"Known as Big Muddy, the Mississippi is the longest"*).
- ***Main***: Factóides, listas e definições.
    - O componente factóide desta tarefa distinguia-se pela exigência de uma resposta exacta em vez de um extracto (e.g. *"the Mississippi"*).
    - As perguntas de listas exigem ainda a construção da resposta a partir de múltiplos

documentos (e.g. *"What Chinese provinces have a McDonald's restaurant?"*).

- As definições exigem respostas mais complexas que não podem ser avaliadas simplesmente como certas ou erradas (e.g. *"What is Ph is Biology?"*). A avaliação baseou-se na divisão dos conceitos em pedaços (*nuggets*) atómicos de informação, devendo as respostas conter pelo menos os *nuggets* essenciais.

Os métodos utilizados para responder a perguntas factuais não têm mudado significativamente. Os sistemas geralmente determinam o tipo de resposta esperado, recolhem documentos ou passagens que podem conter a resposta a partir de palavras-chave da pergunta e termos relacionados, e realizam um *matching* entre essas palavras e as passagens recolhidas para extrair uma resposta.

Para responder com listas, muitos dos concorrentes do TREC'03 usaram o mesmo sistema das perguntas factuais, alterando apenas o número de respostas dadas (problema: determinar o número de respostas adequado). Nas respostas com definições, o importante era encontrar o maior número de *nuggets* de informação. Usaram-se técnicas distintas; por exemplo, *pattern-matching* para encontrar padrões, tais como a presença de aposições nas frases. Tentavam depois eliminar informação redundante, com medidas de sobreposição de palavras ou técnicas de sumarização automática.

Foi a primeira vez que foram integradas perguntas de definições, e a primeira vez que as perguntas de listas tiveram um número significativo de participantes. Os resultados mostraram que estas tarefas apresentam desafios, assim como a sua avaliação. A necessidade de um maior número de perguntas de definições, com vista a solidificar a avaliação, levou a uma reformulação da *QA track* em 2004.

2004

A TREC'04 [33] abordou os 3 tipos de perguntas numa só tarefa: para cada entidade/evento-alvo existia um conjunto de perguntas factuais, outro conjunto de perguntas de listas, e uma pergunta final ("other") pedindo mais informação, equiparável às perguntas de definição da TREC 2003. O conjunto total de perguntas para cada alvo constitui uma série.

Exemplo acerca de um escritor:

- Factóides: data de nascimento/morte, nacionalidade;

- Lista: livros do autor;
- Última pergunta: inclui itens como um dos livros ter ganho o prémio X, ou o autor trabalhar na universidade Y.

Dificuldade adicional: reconhecer/remover informação já dada (um dos aspectos essenciais em RP interactiva). O alvo e as perguntas anteriores formam um contexto para a última pergunta. Não houve grandes novidades em termos tecnológicos (a pergunta "other" foi tratada pelos grupos concorrentes como as perguntas de definições da TREC'03).

2005

A TREC'05 [34] manteve o mesmo esquema, permitindo ainda que os conceitos-alvo fossem também eventos, para além de pessoas, organizações e entidades. Esta mudança foi motivada especialmente pelo facto de os documentos utilizados como fonte de respostas serem constituídos por notícias.

Dado o interesse em examinar o papel das técnicas de recuperação de documentos no apoio a RP, foi exigido a cada participante um *ranking* dos documentos usados para responder a cada questão. Estes dados irão servir de base para comparar as diferentes técnicas de recuperação de documentos.

Outra novidade na TREC'05 foi a abordagem às perguntas relacionais (*relationship questions*), proposta como tarefa opcional. Definiu-se relação como a capacidade de uma entidade influenciar outra, e identificaram-se 8 esferas de influência: financeira, movimento de bens, laços familiares, linhas (*pathways*) de comunicação, laços de organização, co-localização, interesses comuns, e temporal. Foi fornecida aos sistemas uma declaração que definia o contexto para uma questão final sobre um dos tipos de influência. Cada respostas consistiu num conjunto de *nuggets* de informação que proporcionavam as evidências (ou falta delas) para a hipotética relação. Eis um exemplo:

- Questão e contexto:
  - *The analyst is concerned with arms trafficking to Colombian insurgents. Specifically, the analyst would like to know of the different routes used for arms entering Colombia and the entities involved.*
- *Nuggets*:
  - *Weapons are flown from Jordan to Peru and air dropped over southern Columbia*

- *Jordan denied that it was involved in smuggling arms to Columbian guerrillas*
- *Jordan contends that a Peruvian general purchased the rifles and arranged to have them shipped to Columbia via the Amazon River.*
- *Jordan denied that it was involved in smuggling arms to Columbian guerrillas*
- *FARC receives arms shipments from various points including Ecuador and the Pacific and Atlantic coasts.*
- *Entry of arms to Columbia comes from different borders, not only Peru*

2006

Em 2006, a TREC/QA incluiu uma tarefa secundária denominada *complex, interactive Question Answering* [29]. Esta tarefa pretende promover o desenvolvimento de sistemas capazes de processar pedidos de informação consistindo num *template* e numa narrativa (que elabora o que o utilizador procura, providencia contexto, etc).

Exemplo:

- ***Template**: What evidence is there for transport of [cigarettes] from [North Carolina] to [Michigan]?*
- ***Narrative**: The analyst wants to know if there is any evidence that cigarettes are being purchased in North Carolina and then illegally transported and resold in northern states such as Michigan which levy much higher taxes on tobacco.*

A componente de interacção é opcional e deve permitir ao sistema solicitar informação aos utilizadores através de formulários HTML.

ANEXO C

Exemplo de execução do sistema

```
Questão: "quem é edward norton?"
Tipo de questão: parcial "quem", predicativa
Processamento da questão em: 415 ms.


Query (é+edward+norton) enviada para o Google
URL a enviar: http://www.google.pt/search?q=é+edward+norton&lr=lang_pt&hl=pt-PT

Query (é+edward+norton) enviada para o ASKJ
URL a enviar: http://www.ask.com/web?q=é+edward+norton&dm=adv&advl=pt

Query (é+edward+norton) enviada para o Yahoo
URL a enviar: http://search.yahoo.com/search?p=é+edward+norton&fl=1&vl=lang_pt

Query (é+edward+norton) enviada para o MSN
URL a enviar: http://search.msn.com/results.aspx?q=é+edward+norton+language%3Apt



URL:www.geocities.com/Hollywood/Theater/3451/spotlight.html
URL ORIGINAL: www.geocities.com/Hollywood/Theater/3451/spotlight.html
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 1
TIPO: HTML
SNIPPET: Gostoso! - e adjetivos do tipo - tenha certeza de apenas uma coisa: este indivíduo
não é Edward Norton Colocando as cartas na mesa, ele não tem nada demais. É narigudo, tem uns
dentes de ...
------------


URL:pt.wikipedia.org/wiki/Edward_Norton
URL ORIGINAL: pt.wikipedia.org/wiki/Edward_Norton
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 28 KB
RANK: 1
TIPO: null
SNIPPET: Edward Norton (Boston, 18 de Agosto de 1969) é um ator norte-americano. ... O
Wikiquote tem uma coleção de citações de ou sobre: Edward Norton. ...
------------


URL:www.geocities.com/edward_nortononlinewebsite/portugues1.html
URL ORIGINAL: www.geocities.com/edward_nortononlinewebsite/portugues1.html
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 14 KB
RANK: 2
TIPO: HTML
SNIPPET: É proibida a modificação, distribuição, transmissão e re-publição de Edward Norton
Online website, sem autorização prévia por escrito. ...
------------


URL:cinema.yahoo.com.br/perfil/82/fotos/edwardnorton
URL ORIGINAL: cinema.yahoo.com.br/perfil/82/fotos/edwardnorton
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 2
TIPO: null
```

SNIPPET: O Yahoo! Cinema tem a biografia e filmografia de Edward Norton. Veja também fotos, trailers de seus filmes e leia as notícias mais completas e atuais
---

URL:www.geocities.com/edward_nortononlinewebsite/portugues2.html
URL ORIGINAL: www.geocities.com/edward_nortononlinewebsite/portugues2.html
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 36 KB
RANK: 3
TIPO: HTML
SNIPPET: Seu pai, Edward Norton Sr., é advogado e vice-presidente honorário da 'Organização Nacional para Preservação Histórica'. Sua mãe, Robin Norton, ...

URL:adorocinema.cidadeinternet.com.br/filmes/despertar-de-uma-paixao/despertar-de-uma-paixao.asp
URL ORIGINAL: adorocinema.cidadeinternet.com.br/filmes/despertar-de-uma-paixao/despertar-de-uma-paixao.asp
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 3
TIPO: null
SNIPPET: Durante 5 anos a produtora Sara Colleton, o roteirista Ron Nyswaner e o ator Edward Norton trabalharam para que O Despertar de uma Paixão fosse realizado.

URL:www.cineclick.com.br/1000filmes
URL ORIGINAL: www.cineclick.com.br/1000filmes
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 4
TIPO: null
SNIPPET: Brad Pitt e Edward Norton protagonizam este moderno e instigante drama que ganhou o Oscar de Melhores Efeitos Sonoros. ...

URL:epipoca.uol.com.br/gente_detalhes.php?idg=306
URL ORIGINAL: epipoca.uol.com.br/gente_detalhes.php?idg=306
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 26 KB
RANK: 4
TIPO: null
SNIPPET: Edward Norton estreou no cinema interpretando um homem com dupla personalidade em "As Duas Faces de um Crime" e voltou a interpretar alguém com o mesmo ...

URL:ultimosegundo.ig.com.br/materias/cultura/2634001-2634500/2634135/2634135_1.xml
URL ORIGINAL: ultimosegundo.ig.com.br/materias/cultura/2634001-2634500/2634135/2634135_1.xml
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 38 KB
RANK: 5
TIPO: UNSUPPORTED
SNIPPET: NOVA YORK (Reuters) - O ator Edward Norton vem surpreendendo alguns com os papéis mais românticos em filmes de época lançados neste ano, mas ainda não pode ...
---

* Tipo de documento não suportado @ ultimosegundo.ig.com.br/materias/cultura/2634001-2634500/2634135/2634135_1.xml

URL:www.cineplayers.com/perfil.php?id=365
URL ORIGINAL: www.cineplayers.com/perfil.php?id=365
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 19 KB
RANK: 6
TIPO: null
SNIPPET: Sinopse: Derek Vinyard (Edward Norton) era o líder de uma violenta gangue racista, até ser preso e reavaliar seus conceitos. Quando sai, vê que seu irmão ...

```
------------


URL:smilercinemablog.org/?p=1569
URL ORIGINAL: smilercinemablog.org/?p=1569
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 5
TIPO: null
SNIPPET: Ainda mais quando se tem Naomi Watts e Edward Norton para dar paixão a seus
personagens. O Despertar de uma Paixão direção: John Curran.Com Edward Norton, Naomi Watts,
Liev Schreiber, Toby Jones ...
------------


URL:g1.globo.com/Noticias/Cinema/0,,MUL103365-7086,00.html
URL ORIGINAL: g1.globo.com/Noticias/Cinema/0,,MUL103365-7086,00.html
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 42 KB
RANK: 7
TIPO: HTML
SNIPPET: Divulgação Brad Pitt e Edward Norton vão repetir a dobradinha de 'Clube da luta'
(Foto: Divulgação) SAIBA MAIS Brad Pitt e Edward Norton vão atuar juntos ...
------------


URL:www1.folha.uol.com.br/folha/ilustrada/ult90u327850.shtml
URL ORIGINAL: www1.folha.uol.com.br/folha/ilustrada/ult90u327850.shtml
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 29 KB
RANK: 8
TIPO: null
SNIPPET: Brad Pitt e Edward Norton vão atuar juntos novamente no longa-metragem "State of
Play", informou hoje a revista "Variety". ...
------------


* java.io.FileNotFoundException: http://adorocinema.cidadeinternet.com.br/filmes/despertar-de-
uma-paixao/despertar-de-uma-paixao.asp

* Download de www.geocities.com/edward_nortononlinewebsite/portugues1.html concluido.

* Download de www.geocities.com/edward_nortononlinewebsite/portugues2.html concluido.

URL:cinema.ptgate.pt/quem.php?code=137
URL ORIGINAL: cinema.ptgate.pt/quem.php?code=137
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 22 KB
RANK: 9
TIPO: null
SNIPPET: Edward Norton Edward James Norton Jr. data de nascimento 1969-08-18 ... blog links. o
que são blog links? Não há blogs sugeridos sobre Edward Norton. ...
------------


* Download de www.geocities.com/Hollywood/Theater/3451/spotlight.html concluido.

URL:www.odarainternet.com.br/supers/cinema/score.htm
URL ORIGINAL: www.odarainternet.com.br/supers/cinema/score.htm
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 6
TIPO: HTML
SNIPPET: Marlon Brando, Robert de Niro e Edward Norton são ladrões em busca de dinheiro por
motivos diferentes. O mais novo filme de Robert de Niro (de "Jackie Brown") é ...
------------


URL:cineminha.uol.com.br/materia.cfm?id=5049
URL ORIGINAL: cineminha.uol.com.br/materia.cfm?id=5049
MOTOR BUSCA: Google
TAMANHO ESTIMADO: 28 KB
RANK: 10
```

TIPO: null
SNIPPET: Oito anos após a parceria de sucesso em 'Clube da Luta', Edward Norton e Brad Pitt voltarão a dividir as telonas do cinema. A Variety revelou que ambos ...

URL:www.odarainternet.com.br/supers/cinema/frida.htm
URL ORIGINAL: www.odarainternet.com.br/supers/cinema/frida.htm
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 7
TIPO: HTML
SNIPPET: ... Banderas (de "Zorro"), Alfred Molina (de "Magnolia"), Geoffrey Rush (de "Heróis Muito Loucos"), Ashley Judd (de "Beijos Que Matam") e Edward Norton ...

URL:globosat.globo.com/telecine/servicos/artista.asp?id=4395
URL ORIGINAL: globosat.globo.com/telecine/servicos/artista.asp?id=4395
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 8
TIPO: null
SNIPPET: Edward Norton é filho de um advogado que já trabalhou no departamento de acusação federal durante a administração Carter e de uma professora de inglês que faleceu em virtude de um tumor no ...
---

* Download de pt.wikipedia.org/wiki/Edward_Norton concluido.

* Download de www.odarainternet.com.br/supers/cinema/score.htm concluido.

* Download de www.odarainternet.com.br/supers/cinema/frida.htm concluido.

URL:www.focusfilmes.com.br/prd/noticias.php?c=14
URL ORIGINAL: www.focusfilmes.com.br/prd/noticias.php?c=14
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 9
TIPO: null
SNIPPET: A Focus Filmes foi inaugurada em agosto de 2005, pela iniciativa de três sócios apreciadores da ... Magia no 3º Amazonas Film Festival O Ilusionista, filme com Edward Norton, Paul Giamatti e ...
---

* Download de cinema.ptgate.pt/quem.php?code=137 concluido.

URL:www.amce.com.br/frames/sugestoes_detalhe.asp?sut_ID=3&offset=1
URL ORIGINAL: www.amce.com.br/frames/sugestoes_detalhe.asp?sut_ID=3&offset=1
MOTOR BUSCA: MSN
TAMANHO ESTIMADO: -1 KB
RANK: 10
TIPO: null
SNIPPET: A Outra História Americana American History X Edward Furlong e Edward Norton, indicado ao Oscar 119 minutos, 1998 Derek (Edward Norton) precisou de três anos de prisão para descobrir que ...

URL:www.geocities.com/edward_nortononlinewebsite/portugues2.html"%20>Edward%20Norton%20Online%20website%20|%20INTERNATIONAL%20SECTION%20-%20PortuguÃªs%20...%20-%20Translate%20this%20pageWebsite.%20PERFIL.%20FILMOGRAFIA.%20The%20Score.%20Tenha%20FÃ©.%20Clube%20da%20...%20Cartas%20na%20Mesa.%20O%20Povo%20Contra%20Larry%20Flint.%20Todos%20Dizem%20Eu%20te%20Amo.%20As%20Duas%20Faces%20de%20um%20Crime%20...www.geocities.com/edward_nortononlinewebsite/portugues2.html%20-%2035k%20-
URL ORIGINAL: www.geocities.com/edward_nortononlinewebsite/portugues2.html"%20>Edward%20Norton%20Online%20website%20|%20INTERNATIONAL%20SECTION%20-%20PortuguÃªs%20...%20-%20Translate%20this%20pageWebsite.%20PERFIL.%20FILMOGRAFIA.%20The%20Score.%20Tenha%20FÃ©.%20Clube%20da%20...%20Cartas%20na%20Mesa.%20O%20Povo%20Contra%20Larry%20Flint.%20Todos%20Dizem%20Eu%20te%20Amo.%20As%20Duas%20Faces%20de%20um%20Crime%20...www.geocities.com/edward_nortononlinew

ebsite/portugues2.html%20-%2035k%20-
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 35 KB
RANK: 1
TIPO: null
SNIPPET: Website. PERFIL. FILMOGRAFIA. The Score. Tenha FÃ©. Clube da ... Cartas na Mesa. O Povo Contra Larry Flint. Todos Dizem Eu te Amo. As Duas Faces de um Crime ...

URL:pt.wikipedia.org/wiki/Edward_Norton"%20>Edward%20Norton%20-%20WikipÃ©dia%20-%20Translate%20this%20pageEdward%20Norton%20(Boston,%2018%20de%20Agosto%20de%201969)%20Ã©%20um%20ator%20norte-americano.%20...%20O%20avÃ´%20de%20Norton%20fundou%20a%20cidade%20de%20Columbia,%20nos%20EUA,%20e%20Ã©%20conhecido%20como%20o%20...Quick%20Links:
URL ORIGINAL: pt.wikipedia.org/wiki/Edward_Norton"%20>Edward%20Norton%20-%20WikipÃ©dia%20-%20Translate%20this%20pageEdward%20Norton%20(Boston,%2018%20de%20Agosto%20de%201969)%20Ã©%20um%20ator%20norte-americano.%20...%20O%20avÃ´%20de%20Norton%20fundou%20a%20cidade%20de%20Columbia,%20nos%20EUA,%20e%20Ã©%20conhecido%20como%20o%20...Quick%20Links:
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 27 KB
RANK: 2
TIPO: null
SNIPPET: Edward Norton (Boston, 18 de Agosto de 1969) Ã© um ator norte-americano. ... O avÃ´ de Norton fundou a cidade de Columbia, nos EUA, e Ã© conhecido como o ...

URL:www.geocities.com/zoomza/spotlight.html"%20>Zoom%20Zine%20-%20Translate%20this%20page...%20de%20apenas%20uma%20coisa:%20este%20indivÃ-duo%20nÃ£o%20Ã©%20Edward%20Norton%20...%20Verdade%20pura%20e%20cristalina,%20Edward%20Norton%20Ã©%20um%20ator%20perfeito%20e%20nÃ£o%20estÃ¡%20para%20brincadeira.%20...www.geocities.com/zoomza/spotlight.html%20-%2012k%20-
URL ORIGINAL: www.geocities.com/zoomza/spotlight.html"%20>Zoom%20Zine%20-%20Translate%20this%20page...%20de%20apenas%20uma%20coisa:%20este%20indivÃ-duo%20nÃ£o%20Ã©%20Edward%20Norton%20...%20Verdade%20pura%20e%20cristalina,%20Edward%20Norton%20Ã©%20um%20ator%20perfeito%20e%20nÃ£o%20estÃ¡%20para%20brincadeira.%20...www.geocities.com/zoomza/spotlight.html%20-%2012k%20-
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 12 KB
RANK: 3
TIPO: null
SNIPPET: ... de apenas uma coisa: este indivÃduo nÃ£o Ã© Edward Norton ... Verdade pura e cristalina, Edward Norton Ã© um ator perfeito e nÃ£o estÃ¡ para brincadeira. ...
----

* java.io.IOException: Server returned HTTP response code: 403 for URL: http://www.geocities.com/edward_nortononlinewebsite/portugues2.html"%20>Edward%20Norton%20Online%20website%20|%20INTERNATIONAL%20SECTION%20-%20PortuguÃªs%20...%20-%20Translate%20this%20pageWebsite.%20PERFIL.%20FILMOGRAFIA.%20The%20Score.%20Tenha%20FÃ©.%20Clube%20da%20...%20Cartas%20na%20Mesa.%20O%20Povo%20Contra%20Larry%20Flint.%20Todos%20Dizem%20Eu%20te%20Amo.%20As%20Duas%20Faces%20de%20um%20Crime%20...www.geocities.com/edward_nortononlinewebsite/portugues2.html%20-%2035k%20-

URL:www.judao.com.br/cinema/noticia/edward-norton-hulk/"%20>Edward%20Norton%20Ã©%20o%20IncrÃ-vel%20Hulk!%20|%20JUDÃƒO%20-%20Translate%20this%20pageEdward%20Norton%20Ã©%20o%20IncrÃ-vel%20Hulk!%20Segunda-feira,%2016%20de%20Abril%20de%202007.%20THIAGO%20BORBOLLA%20...%20Zak%20Penn%20fala%20sobre%20Edward%20Norton%20como%20Hulk!%20...www.judao.com.br/cinema/noticia/edward-norton-hulk%20-%2048k%20-
URL ORIGINAL: www.judao.com.br/cinema/noticia/edward-norton-hulk/"%20>Edward%20Norton%20Ã©%20o%20IncrÃvel%20Hulk!%20|%20JUDÃƒO%20-%20Translate%20this%20pageEdward%20Norton%20Ã©%20o%20IncrÃvel%20Hulk!%20Segunda-feira,%2016%20de%20Abril%20de%202007.%20THIAGO%20BORBOLLA%20...%20Zak%20Penn%20fala%20sobre%20Edward%20Norton%20como%20Hulk!%20...www.judao.com.br/cinema/noticia/edward-norton-hulk%20-%2048k%20-
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 48 KB
RANK: 4
TIPO: null
SNIPPET: Edward Norton Ã© o IncrÃvel Hulk! Segunda-feira, 16 de Abril de 2007. THIAGO BORBOLLA

... Zak Penn fala sobre Edward Norton como Hulk! ...

URL:ecrangigante.blogspot.com/2006/11/edward-norton-o-grande-o-ilusionista.html"%20>EcranGigante%20-%20Posters,%20Wallpapers%20&%20Trailers%20de%20Filmes.:%20Edward%20Norton%20...%20-%20Translate%20this%20pageEdward%20Norton%20o%20Grande,%20Ã©%20"O%20Ilusionista".%20Quando%20Eisenheim%20(Edward%20Norton)%20comeÃ§a%20a%20executar%20as%20suas%20incriveis%20ilusÃµes%20na%20...ecrangigante.blogspot.com/2006/11/edward-norton-o-grande-o-ilusioni...%20-%2089k%20-
URL ORIGINAL: ecrangigante.blogspot.com/2006/11/edward-norton-o-grande-o-ilusionista.html"%20>EcranGigante%20-%20Posters,%20Wallpapers%20&%20Trailers%20de%20Filmes.:%20Edward%20Norton%20...%20-%20Translate%20this%20pageEdward%20Norton%20o%20Grande,%20Ã©%20"O%20Ilusionista".%20Quando%20Eisenheim%20(Edward%20Norton)%20comeÃ§a%20a%20executar%20as%20suas%20incriveis%20ilusÃµes%20na%20...ecrangigante.blogspot.com/2006/11/edward-norton-o-grande-o-ilusioni...%20-%2089k%20-
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 89 KB
RANK: 5
TIPO: null
SNIPPET: Edward Norton o Grande, Ã© "O Ilusionista". Quando Eisenheim (Edward Norton) comeÃ§a a executar as suas incriveis ilusÃµes na ...

URL:cinetotal.com.br/noticia.php%3fcod=1352"%20>CINETOTAL%20-%20De%20Olho%20no%20Mundo%20Cinematografico%20-%20Translate%20this%20page...%20e%20desenvolvimentos%20intensos%20Nyswaner%20enviou%20o%20roteiro%20para%20Edward%20Norton.%20...%20e%20Kitty%20como%20um%20casal%20Ã©%20admirÃ¡vel,%20principalmente%20a%20maneira%20como%20transcendem%20...cinetotal.com.br/noticia.php?cod=1352%20-%2076k%20-
URL ORIGINAL: cinetotal.com.br/noticia.php%3fcod=1352"%20>CINETOTAL%20-%20De%20Olho%20no%20Mundo%20Cinematografico%20-%20Translate%20this%20page...%20e%20desenvolvimentos%20intensos%20Nyswaner%20enviou%20o%20roteiro%20para%20Edward%20Norton.%20...%20e%20Kitty%20como%20um%20casal%20Ã©%20admirÃ¡vel,%20principalmente%20a%20maneira%20como%20transcendem%20...cinetotal.com.br/noticia.php?cod=1352%20-%2076k%20-
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 76 KB
RANK: 6
TIPO: null
SNIPPET: ... e desenvolvimentos intensos Nyswaner enviou o roteiro para Edward Norton. ... e Kitty como um casal Ã© admirÃ¡vel, principalmente a maneira como transcendem ...

URL:www.oindividuo.com/idiotice/idiota42.htm"%20>Fascismo%20anti-fascista%20-%20Translate%20this%20page...%20irmÃ£o%20mais%20velho%20(interpretado%20por%20Edward%20Norton),%20nazista%20e%20preso%20por%20atirar%20em%20...%20de%20Norton,%20para%20ser%20caracterizado%20como%20nazista,%20Ã©%20mostrado%20discursando%20...www.oindividuo.com/idiotice/idiota42.htm%20-%2018k%20-
URL ORIGINAL: www.oindividuo.com/idiotice/idiota42.htm"%20>Fascismo%20anti-fascista%20-%20Translate%20this%20page...%20irmÃ£o%20mais%20velho%20(interpretado%20por%20Edward%20Norton),%20nazista%20e%20preso%20por%20atirar%20em%20...%20de%20Norton,%20para%20ser%20caracterizado%20como%20nazista,%20Ã©%20mostrado%20discursando%20...www.oindividuo.com/idiotice/idiota42.htm%20-%2018k%20-
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 18 KB
RANK: 7
TIPO: null
SNIPPET: ... irmÃ£o mais velho (interpretado por Edward Norton), nazista e preso por atirar em ... de Norton, para ser caracterizado como nazista, Ã© mostrado discursando ...
---

* java.io.FileNotFoundException: http://www.geocities.com/zoomza/spotlight.html"%20>Zoom%20Zine%20-%20Translate%20this%20page...%20de%20apenas%20uma%20coisa:%20este%20indivÃ-duo%20nÃ£o%20Ã©%20Edward%20Norton%20...%20Verdade%20pura%20e%20cristalina,%20Edward%20Norton%20Ã©%20um%20ator%20perfeito%20e%20nÃ£o%20estÃ¡%20para%20brincadeira.%20...www.geocities.com/zoomza/spotlight.html%20-%2012k%20-

* Download de smilercinemablog.org/?p=1569 concluido.

* java.io.IOException: Server returned HTTP response code: 400 for URL:
http://ecrangigante.blogspot.com/2006/11/edward-norton-o-grande-o-
ilusionista.html"%20>EcranGigante%20-
%20Posters,%20Wallpapers%20&%20Trailers%20de%20Filmes.:%20Edward%20Norton%20...%20-
%20Translate%20this%20pageEdward%20Norton%20o%20Grande,%20Ã©%20"O%20Ilusionista".%20Quando%20E
isenheim%20(Edward%20Norton)%20comeÃ§a%20a%20executar%20as%20suas%20incriveis%20ilusÃµes%20na%
20...ecrangigante.blogspot.com/2006/11/edward-norton-o-grande-o-ilusioni...%20-%2089k%20-

URL:lestrange.wordpress.com/"%20>LeStrange%20-
%20Translate%20this%20page...%20Liv%20Tyler%20como%20Betty%20Ross%20e%20Edward%20Norton%20como
%20Bruce%20Banner,%20parece%20mais%20fÃ¡cil%20...%20Tenho%20a%20dizer%20que%20o%20senhor%20Nor
ton%20Ã©%20dos%20meus%20actores%20preferidos%20desde%20a%20...lestrange.wordpress.com%20-
%2048k%20-
URL ORIGINAL: lestrange.wordpress.com/"%20>LeStrange%20-
%20Translate%20this%20page...%20Liv%20Tyler%20como%20Betty%20Ross%20e%20Edward%20Norton%20como
%20Bruce%20Banner,%20parece%20mais%20fÃ¡cil%20...%20Tenho%20a%20dizer%20que%20o%20senhor%20Nor
ton%20Ã©%20dos%20meus%20actores%20preferidos%20desde%20a%20...lestrange.wordpress.com%20-
%2048k%20-
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 48 KB
RANK: 8
TIPO: null
SNIPPET: ... Liv Tyler como Betty Ross e Edward Norton como Bruce Banner, parece mais fÃ¡cil
... Tenho a dizer que o senhor Norton Ã© dos meus actores preferidos desde a ...

URL:www.omelete.com.br/cine/100007418/Interferencia_de_Edward_Norton_no_roteiro_de_Hulk_e_expl
icada.aspx"%20>Omelete%20-
%20InterferÃªncia%20de%20Edward%20Norton%20no%20roteiro%20de%20Hulk%20Ã©%20explicada%20-
%20Translate%20this%20pageInterferência%20de%20Edward%20Norton%20no%20roteiro%20de%20Hulk%20é%
20explicada%20...%20Hulk%20foi%20reescrito%20por%20Edward%20Norton,%20ator%20recÃ©m-
contratado%20para%20viver%20Bruce%20Banner.%20...omelete.com.br/cine/100007418/Interferencia_d
e_Edward_Norton_no_rot...%20-%2045k%20-
URL ORIGINAL:
www.omelete.com.br/cine/100007418/Interferencia_de_Edward_Norton_no_roteiro_de_Hulk_e_explicad
a.aspx"%20>Omelete%20-
%20InterferÃªncia%20de%20Edward%20Norton%20no%20roteiro%20de%20Hulk%20Ã©%20explicada%20-
%20Translate%20this%20pageInterferência%20de%20Edward%20Norton%20no%20roteiro%20de%20Hulk%20é%
20explicada%20...%20Hulk%20foi%20reescrito%20por%20Edward%20Norton,%20ator%20recÃ©m-
contratado%20para%20viver%20Bruce%20Banner.%20...omelete.com.br/cine/100007418/Interferencia_d
e_Edward_Norton_no_rot...%20-%2045k%20-
MOTOR BUSCA: Yahoo
TAMANHO ESTIMADO: 45 KB
RANK: 9
TIPO: null
SNIPPET: Interferência de Edward Norton no roteiro de Hulk é explicada ... Hulk foi reescrito
por Edward Norton, ator recÃ©m-contratado para viver Bruce Banner. ...
----

* java.io.FileNotFoundException:
http://www.omelete.com.br/cine/100007418/Interferencia_de_Edward_Norton_no_roteiro_de_Hulk_e_e
xplicada.aspx"%20>Omelete%20-
%20InterferÃªncia%20de%20Edward%20Norton%20no%20roteiro%20de%20Hulk%20Ã©%20explicada%20-
%20Translate%20this%20pageInterferência%20de%20Edward%20Norton%20no%20roteiro%20de%20Hulk%20é%
20explicada%20...%20Hulk%20foi%20reescrito%20por%20Edward%20Norton,%20ator%20recÃ©m-
contratado%20para%20viver%20Bruce%20Banner.%20...omelete.com.br/cine/100007418/Interferencia_d
e_Edward_Norton_no_rot...%20-%2045k%20-

* Download de cinema.yahoo.com.br/perfil/82/fotos/edwardnorton concluido.

* Download de pt.wikipedia.org/wiki/Edward_Norton"%20>Edward%20Norton%20-%20WikipÃ©dia%20-
%20Translate%20this%20pageEdward%20Norton%20(Boston,%2018%20de%20Agosto%20de%201969)%20Ã©%20um
%20ator%20norte-americano.%20...%20O%20avÃ
´%20de%20Norton%20fundou%20a%20cidade%20de%20Columbia,%20nos%20EUA,%20e%20Ã©%20conhecido%20com
o%20o%20...Quick%20Links: concluido.

* java.io.IOException: Server returned HTTP response code: 403 for URL:
http://cinetotal.com.br/noticia.php%3fcod=1352"%20>CINETOTAL%20-
%20De%20Olho%20no%20Mundo%20Cinematografico%20-
%20Translate%20this%20page...%20e%20desenvolvimentos%20intensos%20Nyswaner%20enviou%20o%20rote

```
iro%20para%20Edward%20Norton.%20...%20e%20Kitty%20como%20um%20casal%20Ã©%20admirÃ¡vel,%20princ
ipalmente%20a%20maneira%20como%20transcendem%20...cinetotal.com.br/noticia.php?cod=1352%20-
%2076k%20-

* Download de www.cineplayers.com/perfil.php?id=365 concluido.

* Download de g1.globo.com/Noticias/Cinema/0,,MUL103365-7086,00.html concluido.

* Download de www.focusfilmes.com.br/prd/noticias.php?c=14 concluido.

* Download de epipoca.uol.com.br/gente_detalhes.php?idg=306 concluido.

* Download de www.amce.com.br/frames/sugestoes_detalhe.asp?sut_ID=3&offset=1 concluido.

* java.io.FileNotFoundException: http://lestrange.wordpress.com/"%20>LeStrange%20-
%20Translate%20this%20page...%20Liv%20Tyler%20como%20Betty%20Ross%20e%20Edward%20Norton%20como
%20Bruce%20Banner,%20parece%20mais%20fÃ¡cil%20...%20Tenho%20a%20dizer%20que%20o%20senhor%20Nor
ton%20Ã©%20dos%20meus%20actores%20preferidos%20desde%20a%20...lestrange.wordpress.com%20-
%2048k%20-

* java.io.FileNotFoundException: http://www.judao.com.br/cinema/noticia/edward-norton-
hulk/"%20>Edward%20Norton%20Ã©%20o%20IncrÃvel%20Hulk!%20|%20JUDÃƒO%20-
%20Translate%20this%20pageEdward%20Norton%20Ã©%20o%20IncrÃvel%20Hulk!%20Segunda-
feira,%2016%20de%20Abril%20de%202007.%20THIAGO%20BORBOLLA%20...%20Zak%20Penn%20fala%20sobre%20
Edward%20Norton%20como%20Hulk!%20...www.judao.com.br/cinema/noticia/edward-norton-hulk%20-
%2048k%20-

* Download de globosat.globo.com/telecine/servicos/artista.asp?id=4395 concluido.

* Download de www.cineclick.com.br/1000filmes concluido.

* java.io.FileNotFoundException:
http://www.oindividuo.com/idiotice/idiota42.htm"%20>Fascismo%20anti-fascista%20-
%20Translate%20this%20page...%20irmÃ£o%20mais%20velho%20(interpretado%20por%20Edward%20Norton)
,%20nazista%20e%20preso%20por%20atirar%20em%20...%20de%20Norton,%20para%20ser%20caracterizado%
20como%20nazista,%20Ã©%20mostrado%20discursando%20...www.oindividuo.com/idiotice/idiota42.htm%
20-%2018k%20-

* Download de www1.folha.uol.com.br/folha/ilustrada/ult90u327850.shtml concluido.

* Download de cineminha.uol.com.br/materia.cfm?id=5049 concluido.


Frases candidatas: 90
Frases inuteis: 1082



----------------------
RESPOSTA #1 (valor 6)
----------------------

Resposta exacta: ator perfeito
Resposta exacta: cristalina
Frase completa: Verdade pura e cristalina, Edward Norton é um ator perfeito e não está para
brincadeira.

Data: Mon Nov 15 19:55:37 WET 1999
Documento: www.geocities.com/Hollywood/Theater/3451/spotlight.html
Motor: MSN (rank 1)

Busca e parse de resultado: 514 ms.
Download e parse de documento: 492 ms.
Extraccao de resposta: 258 ms.
Tempo total: 1679 ms.

----------------------
RESPOSTA #2 (valor 5)
----------------------

Resposta exacta: filho
Frase completa: Edward Norton é filho de um advogado que já trabalhou no departamento de
acusação federal durante a administração Carter e de uma professora de inglês que faleceu em
```

virtude de um tumor no cérebro.

Documento: globosat.globo.com/telecine/servicos/artista.asp?id=4395
Motor: MSN (rank 8)

Busca e parse de resultado: 890 ms.
Download e parse de documento: 1954 ms.
Extraccao de resposta: 11 ms.
Tempo total: 3270 ms.

RESPOSTA #3 (valor 4)

Frase: Gostoso! - e adjetivos do tipo - tenha certeza de apenas uma coisa: este indivíduo não é Edward Norton

Data: Mon Nov 15 19:55:37 WET 1999
Documento: www.geocities.com/Hollywood/Theater/3451/spotlight.html
Motor: MSN (rank 1)

Busca e parse de resultado: 514 ms.
Download e parse de documento: 492 ms.
Extraccao de resposta: 258 ms.
Tempo total: 1679 ms.

RESPOSTA #4 (valor 4)

Frase: Incrível Hulk - Edward Norton é anunciado para o papel de Bruce Banner em novo filme (16/04/2007 17:27)

Documento: epipoca.uol.com.br/gente_detalhes.php?idg=306
Motor: Google (rank 4)

Busca e parse de resultado: 561 ms.
Download e parse de documento: 2207 ms.
Extraccao de resposta: 41 ms.
Tempo total: 3224 ms.

RESPOSTA #5 (valor 3)

Frase: Edward Norton (Boston, 18 de Agosto de 1969) é um ator norte-americano.

Data: Thu Sep 06 03:01:37 WEST 2007
Documento: pt.wikipedia.org/wiki/Edward_Norton
Motor: Google (rank 1)

Busca e parse de resultado: 524 ms.
Download e parse de documento: 572 ms.
Extraccao de resposta: 187 ms.
Tempo total: 1698 ms.

***
Respostas redundantes descartadas: 0
FIM: 8920 ms.

ANEXO D

Perguntas "Quem...?" predicativas usadas nas avaliações

- Quem é Joe Satriani?
- Quem é o primeiro-ministro da Macedónia?
- Quem era Mpinga Kassenda?
- Quem foi Charles Bukowski?
- Quem foi Emiliano Zapata?
- Quem é Michel Noir?
- Quem é Eudald Carbonell?
- Quem é a viúva de John Lennon?
- Quem foi o sucessor de Kim Il Sung?
- Quem é Flavio Briatore?
- Quem é Brian Tobin?
- Quem é Sergio Balanzino?
- Quem é Josef Olesky?
- Quem é o presidente do Banco Mundial?
- Quem é o governador do Banco de Inglaterra?
- Quem é Yigal Amir?
- Quem é o Ministro da Defesa holandês?
- Quem é Wim Kok?
- Quem era Edwin Hubble?
- Quem é Chun Doo Hwan?
- Quem é Joe Slovo?
- Quem é o primeiro-ministro de Singapura?
- Quem é Johannes van Damme?
- Quem é o presidente do Peru?
- Quem era Claretta Petacci?

- Quem é Carolos Papoulias?
- Quem é Denis Echard?
- Quem é Fidel Ramos?
- Quem é Joaquin Navarro-Valls?
- Quem é Jáder Barbalho?
- Quem é Tenzin Gyatso?
- Quem é Vladimir Dezhurov?
- Quem é presidente do COI desde 1980?
- Quem era o ditador cubano antes da revolução?
- Quem era a amante de Mussolini?
- Quem foi o primeiro homem em órbita?
- Quem é o advogado de Andreotti?
- Quem é o director geral da AIEA?
- Quem é o líder do Partido Popular Italiano?
- Quem é o patriarca de Alexandria?
- Quem é o prefeito de Lisboa?
- Quem é o presidente da Macedónia?
- Quem é o primeiro-ministro grego?
- Quem é o único atleta português nos Jogos Olímpicos de Lillehammer?
- Quem foi ministro da Justiça de Itália entre 1945 e 1946?
- Quem foi primeiro-ministro de França durante o regime de Vichy?
- Quem é o recordista mundial do salto à vara?
- Quem é a "diva dos pés descalços"?
- Quem é o secretário-geral do PCP?
- Quem é o Presidente da Câmara de Lisboa?
- Quem é o Presidente da Câmara de Lamego?
- Quem é o embaixador de Portugal em França?
- Quem foi o primeiro presidente dos Estados Unidos?
- Quem é o ministro-presidente da Renânia-Palatinado?
- Quem foi o último governador de Timor Leste?
- Quem era o marido de Vieira da Silva?
- Quem é o capitão do FC Porto?
- Quem é o imã da mesquita de Lisboa?

- Quem é a ministra sueca do ambiente?
- Quem é o padroeiro de Penafiel?
- Quem é Leonor Beleza?
- Quem é Arnold Ruutel?
- Quem é Wim Duisenberg?
- Quem é Rocha Vieira?
- Quem é Guilherme da Fonseca?
- Quem é Fernando Gomes?
- Quem é Valentina Terechkova?
- Quem é Jorge Amado?

Referências

[1] Amaral, C.; Laurent, D.; Martins, A.; Mendes, A.; Pinto, C. (2004) "Design and Implementation of a Semantic Search Engine for Portuguese". *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, vol.1, 247–250.

[2] Amaral, C.; Figueira, H.; Martins A.; Mendes, A.; Mendes, P.; Pinto, C. (2005) "Priberam's question answering system for Portuguese" *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop.*

[3] Branco, A. e Silva, J. (2004) "Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers". *Proceedings of the 4th Language Resources and Evaluating Conference,* 507-510.

[4] Breck, E.J.; Burger, J.D.; Ferro, L.; Hirschman, L.; House, D.; Light, M.; Mani, I. (2000) "How to Evaluate Your Question Answering System Every Day ... and Still Get Real Work Done" *Proceedings of the Second International Conference on Language Resources and Evaluation*.

[5] Brill, E. (2003) "Processing Natural Language without Natural Language Processing" *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics,* 360-369.

[6] Carbonell, J.; Harman, D.; Hovy, E.; Maiorano, S.; Prange, J. e Sparck-Jones, K.. (2000) "Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization" Disponível em: http://www-nlpir.nist.gov/projects/duc/papers/Final-Vision-Paper-v1a.pdf

[7] Costa, L.F. (2006) "Esfinge – a Question Answering System in the Web using the Web" *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics.*

[8] Esfinge (SRP) – http://www.linguateca.pt/Esfinge/

[9] Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press.*

[10] Green, B.F., Wolf, A.K., Chomsky, C. and Laughery, K. (1961) "BASEBALL: An automatic question answerer" *Proceedings Western Joint Computer Conference* 19: 219-224.

[11] Harabagiu, S. e Moldovan, D. (2003) "Question Answering" In Mitkov, R. *The Oxford Handbook of Computational Linguistics*, 31: 560-580. Oxford University Press.

[12] Hirschman, L. e Gaizauskas, R. (2001) "Natural language question answering: the view from here", *Natural Language Engineering* 7: 275-300. Cambridge University Press.

[13] Infopédia [em linha]. Porto Editora 2003-2007. Disponível em http://www.infopedia.pt

[14] Kamp,H. e Reyle, U. (1993) *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Kluwer Academic Dordrecht.

[15] Kilgariff, A. (2007) *"Googleology is Bad Science" Computational Linguistics*, vol. 33 (1), 147-151.

[16] Kupiec, J. (1993) "MURAX: a robust linguistic approach for question answering using an online encyclopedia" *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 181-90.

[17] Lehnert, W. (1978) *The Process of Question Answering*. Hilssdale, NJ: Lawrence Erlbaum Associates.

[18] Lin, J. (2005) "Evaluation of resources for question answering evaluation" *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 392-399.

[19] Magnini, B.; Negri, M.; Prevete, R.; Tanev, H. (2002) "Towards Automatic Evaluation of Question/Answering Systems" *Proceedings of the Third International Conference on Language Resources and Evaluation*

[20] Magnini, B.; Giampiccolo, D.; Forner, P.; Ayache, C.; Osenova, P.; Peñas, A.; Jijkoun, V.; Sacaleanu, B.; Rocha, P. e Sutcliffe, R. (2006) "Overview of the CLEF 2006 multilingual

question answering track". *Working Notes for the CLEF 2006 Workshop*.

[21] Mateus, M.; Brito, A.; Duarte, I.; Faria, I. et al. (2003) *Gramática da Língua Portuguesa*. Caminho.

[22] Mollá, D. e Vicedo, J.L. (2007) "Question Answering in Restricted Domains: An Overview" *Computational Linguistics*, vol. 33 (1), 41-61.

[23] Quaresma, P. e Rodrigues, I. (2005) "A logic programming based approach to the QA@CLEF05 track" *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop.*

[24] Quaresma, P. e Rodrigues, I. (2005) "A question-answering system for juridical documents" *Proceedings of the 10th International Conference on Artificial Intelligence and Law.* 256-257.

[25] dos Santos, C.T.; Quaresma, P.; Rodrigues, I. e Vieira, R. (2006) "A Multi-agent Approach to Question Answering" *Proceedings of the 7th International Workshop on Computational Processing of the Portuguese Language (PROPOR 2006)*, Lecture Notes in Artificial Inteligence, núm. 3960, 131-139.

[26] Schank, R. (1972) "Conceptual Dependency: A Theory of Natural Language Understanding". *Cognitive Psychology*, vol. 3 (4), 562-631.

[27] Silva, J. (2007) *Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization*. Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa.

[28] Simmons, R.F. (1965) "Answering English questions by computer: A survey" *Communications Association for Computing Machinery (ACM)* 8(1): 53-70.

[29] TREC 2006 ciQA Task Homepage. http://www.umiacs.umd.edu/~jimmylin/ciqa/

[30] Voorhees, E. e Tice, D. (2000) "Building a Question Answering Test Collection" *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 200-207.

[31] Voorhees, E. e Tice, D. (2000) "Implementing a Question Answering Evaluation" *Proceedings of LREC'2000 Workshop on Using Evaluation within HLT Programs: Results*

and Trends.

[32] Voorhees, E. (2003) "Overview of the TREC 2003 Question Answering Track". TREC 2003.

[33] Voorhees, E. (2004) "Overview of the TREC 2004 Question Answering Track". TREC 2004.

[34] Voorhees, E. (2005) "Overview of TREC 2005" *The Fourteenth Text REtrieval Conference Proceedings.*

[35] Voorhees, E. (2005) "Overview of the TREC 2005 Question Answering Track". *The Fourteenth Text REtrieval Conference Proceedings.*

[36] Wikipédia (entrada sobre resposta a perguntas) – http://en.wikipedia.org/wiki/Question_answering

[37] Winograd, T. (1977) "Five lectures on artificial intelligence" In Zampolli, A. *Linguistic Structures Processing*, vol.5 de *Fundamental Studies in Computer Science*, 521-569. North Holland.

[38] Woods, W.A. (1977) "Lunar rocks in natural English: Explorations in natural language question answering. In Zampolli, A. *Linguistic Structures Processing*, vol.5 of *Fundamental Studies in Computer Science*, 521-569. North Holland.

[39] Zheng, Z. (2002) "AnswerBus Question Answering System". Human Language Technology Conference 2002.

Glossário de Siglas

CLEF	*Cross-Language Evaluation Forum*
CSS	*Cascade Style Sheets* (linguagem de descrição da apresentação de documentos escritos em linguagens como XML e HTML)
DI-FCUL	Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa
EI	Extracção de Informação
FCT	Fundação para a Ciência e Tecnologia
HTML	*Hypertext Markup Language* (a linguagem predominante para escrita de páginas *web*)
I&D	Investigação e Desenvolvimento
ILN	Interacção em Linguagem Natural
NLX	Grupo de Fala e Linguagem Natural do DI-FCUL
PLN	Processamento de Linguagem Natural
QA	*Question Answering* (Inglês: Resposta-a-Perguntas)
RI	Recuperação de Informação
RP	Resposta-a-Perguntas
SN	Sintagma nominal
SRP	Sistema de Resposta-a-Perguntas
TREC/QA	*Text Retrieval Conference: Question Answering track*
URL	*Uniform Resource Locator* (usado no sentido de: endereço de Internet)
XML	*Extensible Markup Language* (uma linguagem de propósito geral, configurável, que combina texto e informação sobre o texto)