UNIVERSIDADE DE ÉVORA

Mestrado em Engenharia Informática

Um Sistema Pergunta/Resposta para a Língua Portuguesa

Pedro Dinis Loureiro Salgueiro

orientador: *Prof. Doutor Paulo Quaresma*

Dezembro de 2004

Esta dissertação não inclui as críticas e sugestões feitas pelo júri.

UNIVERSIDADE DE ÉVORA

Mestrado em Engenharia Informática

Um Sistema Pergunta/Resposta para a Língua Portuguesa

Pedro Dinis Loureiro Salgueiro

orientador: *Prof. Doutor Paulo Quaresma*

Dezembro de 2004

Esta dissertação não inclui as críticas e sugestões feitas pelo júri.

Errata

Na página 13, linha 17, deve substituir-se "(Zekke & Mooney 1996)" por "(Zelle & Mooney 1996)".

Na página 10, linha 1, deve substituir-se "estado da da área" por "estado da arte da área".

Na página 10, linha 28, deve substituir-se "reponderem" por "responderem".

Na página 18, linha 23, deve substituir-se "caracterizador" por "caracterizados".

Na página 19, linha 18, deve substituir-se "consgui-se" por "conseguisse".

Na página 26, linha 9, deve substituir-se "(Moklovan et a1.1999)" por "(Moldovan et al.1999)".

Na página 28, linha 23, deve substituir-se "passage retrievel" por "passage retrieval".

Na página 32, linha 29, deve substituir-se "Sintatic" por "Syntatic".

Na página 35, linha 26, deve substituir-se "existência informação" por "existência de informação".

Na página 37, linha 24, deve substituir-se "palavra substituída" por "palavra for substituída".

Na página 40, linha 14, deve substituir-se "first" "India" por "first" "president" "India".

Na página 51, linha 14, deve substituir-se "cardo" por "cargo".

Na página 57, linha 3, deve substituir-se "perguntas e está por "perguntas e respostas está".

Na página 58, linha 12, deve substituir-se "à feita" por "é feita".

Na página 59, linha 10, deve substituir-se "Língua" por "Lingua Portuguesa".

Na página 67, linha 16, deve substituir-se "desenvolvido o VISL" por "desenvolvido no VISL".

Na página 77, linha 20, deve substituir-se "resposta" por "respostas".

Na página 82, linha 7, deve substituir-se "este sistema de as respostas" por "as respostas".

Na página 100, linha 8, deve substituir-se "pode pelos" por "pode ver pelos".

Na bibliografia devem ser acresentadas as seguintes referências:

Zelle J. M. Using Inductive Logic Programming to Automate the Construction of Natural Languafe Parsers. PhD thesis, Department of Computer Sciences, University of Texas, Austin, TX. Also appears as Artificial Intelligence Laboratory Technical, 1995.

Zelle J. M. and Mooney R. J. Learning to parse database queries using inductive logic programing. pages 1050 1055, Menlo Park CA, 1996. Thirteenth National Conference on Artificial Inteligence, AAAI Press.

Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Richard Goodrum, Roxana G1rju, and Vasile Rus. Lasso: a tool for surfing the answer net. In Proceedings of TREC-8, 1999.

Christiane Fellbaum. WordNet - An Electronic Lexical Database. MIT Press, 1998.

Jimmy Lin and Boris Katz. Question answering techniques for the world wide web. In EACL-2003 Tutorial, 2003.

Jimmy Lin, Aaron Fernandes, Boris Katz, Gregory Marton, and Stefanie Tel-

lex. Extracting answers from the web using knowledge annotation and knowledge mining techniques. In Proceedings of the Eleventh Text REtrieval Conference (TREC 2002), 2002.

Jimmy Lin and Boris Katz. Question answering from theweb using knowledge annotation and knowledge mining techniques. In Proceedings of The Twelfth International Conference on Information and Knowledge Management (CIKM 2003), 2003.

Ian A. Witten, Alistair Mo at, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann, Los Altos, CA, 2nd edition, 1999.

C. Kwok, O. Etzioni, and D. Weld. Scaling question answering to theweb. In Proceedings of WWW10, Hong Kong, 2001.

Eric Brill. Processing natural language without natural language processing. In A. Gelbukh (ed.), CICLing 2003, pages 360 9, LNCS 2588, Springer-Verlag Berlin Heidelberg, 2003.

Prefácio

Este documento contém uma dissertação intitulada *"Um sistema de pergunta / resposta para a língua Portuguesa"*, um trabalho do aluno *Pedro Dinis Loureiro Salgueiro*[1], estudante de Mestrado em Engenharia Informática na Universidade de Évora.

O orientador deste trabalho é o Professor Doutor Paulo Quaresma[2], do Departamento de Informática da Universidade de Évora.

O autor do trabalho é licenciado em Engenharia Informática, pela Universidade de Évora. A presente dissertação foi entregue em Novembro de 2003.

[1] pds@di.uevora.pt
[2] pq@di.uevora.pt

Agradecimentos

Quero aqui agradecer a um conjunto de pessoas que me ajudaram a desenvolver o trabalho aqui apresentado, todas elas parcialmente responsáveis pelo seu conteúdo final.

Primeiramente, quero expressar os meus sinceros agradecimentos ao Professor Paulo Quaresma pelo seu incentivo, acompanhamento e disponibilidade sobre a evolução deste trabalho. Agradeço também à professora Irene Rodrigues pelas produtivas discussões que tivemos sobre PLN que tanto ajudaram neste trabalho.

Agradeço também a toda a minha família pelo grande apoio que me deu ao longo deste trabalho.

Quero também agradecer a todos os meus colegas e amigos que não poderia aqui nomear e que me apoiaram ao longo deste trabalho, em projectos académicos e em outros projectos.

Sumário

Nos dias de hoje existe uma cada vez maior necessidade de informação num mundo cada vez mais rodeado por dados. A necessidade de encontrar informação relevante no meio da grande imensidão disponível de dados torna-se cada vez mais importante, chegando mesmo a tornar-se um factor essencial e fulcral em muitas áreas.

Tendo em consideração esta grande necessidade de obter informação de boa qualidade, os sistemas de recuperação de informação tentam fazer esta recuperação através de um vasto conjunto de dados. Estes sistemas tentam de alguma forma facilitar aos utilizadores a pesquisa de informação sobre um tema, apresentando para isso uma lista de documentos que possam conter informação relevante para o utilizador. O uso destes sistemas tornou-se bastante útil nos últimos anos quando se pretende obter informação sobre algum tema, no entanto mostram algumas falhas pois os utilizadores têm que consultar os vários documentos recuperados para encontrar a informação pretendida.

Como resposta a este problema surgiram os sistemas de perguntas e respostas que recebem dos utilizadores uma pergunta em língua natural e tentam responder de uma forma simples e directa à pergunta. Estes sistemas tornaram-se assim bastante mais úteis que os sistemas de recuperação de informação tradicionais, facilitando ainda mais a grande necessidade de encontrar informação relevante verifica nos dias de hoje.

Esta dissertação estuda uma abordagem e a implementação de um sistema de perguntas e respostas para a língua Portuguesa que tenta responder de uma forma simples e directa a perguntas feitas em língua Portuguesa através de pesquisas em colecções de documentos escritos também em língua Portuguesa.

Abstract

A question answer system
to the
Portuguese language.

Nowadays there is a huge necessity of information in a world where data is produced all over the place. The necessity to find good information in the vast collection data existing all over the world is becoming more important, being in some areas an essential and fulcral factor.

Having in consideration this great necessity to get information with good quality, the information retrieval systems try to retrieve from a vast collection of data important information to the user. These systems try to ease the users to research information on a subject, presenting a list of relevant documents to them can contain relevant information. The use of these systems became very useful in the last years to research information on a subject, however, they show some imperfections, obligating the users to consult the retrieved documents in order to find the intended information.

As a reply to this problem the question and answer systems receive from the users a question in natural language and try to answer that question in a simple and direct way. These systems had become more useful than the traditional information retrieval systems because they make things easier. Users can now find good and quality information.

This work studies an approach and it's implementation of a question and answer system for the Portuguese that tries to answer in a simple and direct form to the questions made in Portuguese through a research in document collections also written in Portuguese language.

Conteúdo

Lista de Figuras

Capítulo 1

Introdução

Esta dissertação descreve um trabalho de investigação numa área especifica de recuperação de informação (*Information Retrieval*): os sistemas de perguntas e respostas (*Question Answer Systems*), sendo estudados de uma forma muito geral alguns sistemas de perguntas e respostas para várias línguas, algumas técnicas e teorias usadas neste tipo de sistemas. O principal foco deste de estudo trabalho é um sistema de perguntas e respostas para língua Portuguesa.

1.1 Motivação

A necessidade de obter informação relevante hoje em dia é cada vez mais importante, sendo um factor critico em muitas áreas. Uma das formas muito comuns de obter informação é através de sistemas de recuperação de informação.

Estes sistemas têm como tarefas fazer a representação, armazenamento, organização e garantir um acesso fácil e rápido à informação

Os sistemas de bases de dados podem ser considerados como sistemas de recuperação de informação, no entanto são apenas sistemas de recuperação de dados que apenas tentam encontrar dados que contenham todas as palavras que pertencem à interrogação feita pelo utilizador. Este tipo de pesquisa não é suficiente para os sistemas de recuperação de informação pois os dados podem estar feitos de uma forma não regular, conter erros e serem muito ambíguos.

Uma grande diferença entre os dois tipos de sistemas é que os dados dos sistemas de recuperação de dados são bem estruturados, enquanto que dos sistemas de recuperação de informação não têm qualquer tipo de estrutura e são muito ambíguos pois normalmente são textos em língua natural.

Os sistemas de recuperação de dados apenas conseguem fornecer os dados existentes, não conseguindo fazer a extracção de informação para apresentar ao utilizador

Para se conseguir fornecer informação ao utilizador sobre um dado assunto é necessário que o sistema consiga fazer a interpretação dos dados e classifica-los de acordo com a relevância que estes têm para a interrogação feita pelo utilizador. Esta interpretação pode passar pela extracção de informação sintáctico/semântica e fazer uma unificação desta informação com a interrogação do utilizador ao sistema.

A relevância que os dados têm para a interrogação é um dos tópicos mais importantes para os sistemas de recuperação de informação pois estes tentam encontrar o maior número possível de documentos muito relevantes e um número mínimo de documentos muito pouco relevantes para a interrogação do utilizador.

Os sistemas de recuperação de informação mudaram muito desde que estes apareceram, permitindo actualmente fazer a modelação, classificação e categorização dos documentos, têm interfaces gráficos, permitem fazer a visualização dos dados e filtrar os dados.

Embora estes sistemas tenham sofrido uma grande evolução e tenha havido um grande aumento do uso de computadores pessoais ao longo dos anos, apenas com o aparecimento da WEB no inicio da década de 1990 é que estes sistemas começaram a ser usados com mais frequência.

A Web está a tornar-se a maior base de conhecimento universal, permitindo como nunca a partilha de ideias e informação a uma escala nunca antes vista. Este grande sucesso deve-se a um interface que pouco muda independentemente do ambiente computacional do utilizador. Um utilizador pode criar os seus documentos e partilhá-los através da Web praticamente a custo zero, podendo este ser acedido em qualquer lugar a qualquer hora. Todos estes factos ajudam para o grande crescimento da Web e para o seu grande sucesso.

Apesar do grande sucesso da Web, esta tem alguns problemas devido à imensa quantidade de informação disponível.

A procura de informação na Web pode tornar-se uma tarefa difícil, levando o utilizador a navegar através da vasta informação existente na Web (usando *links*) à procura de informação que seja interessante para o utilizador. Esta navegação pela Web pode-se tornar muito pouco eficiente pois o espaço de procura da Web é imenso.

O principal problema da Web é a falta de uma estrutura bem definida , levando a que a informação existente na Web seja muitas vezes mal estruturada.

A dificuldade de encontrar informação na Web tornou os sistemas de recuperação de informação em sistemas bastante usados por milhões de utilizadores, conseguindo-se assim obter bons resultados através do uso das suas técnicas de pesquisa.

Este grande uso dos sistemas de recuperação na Web levou muitos investigadores a começarem a trabalhar mais nestes sistemas fazendo com que tenha havido uma grande evolução, e tornando os sistemas bastante mais fiáveis.

Os sistemas de recuperação de informação são os sistemas ideais para encontrar informação na Web, uma vez que os dados disponíveis na Web são muito pouco estruturados, ambíguos e na maior parte das vezes são textos em língua natural.

O Google[1], um dos maiores motores de busca consegue fazer pesquisa de informação em 8.058.044.651 documentos disponíveis na Web. Este número de documentos foi visto no dia 19 de Novembro de 2004 estando este número sempre a crescer. Estes valores mostram que a Web e consequentemente os sistemas de recuperação de informação têm cada vez mais importância na vida das pessoas e cada vez são mais usados.

Os sistemas de recuperação têm vindo a evoluir em várias formas, evoluindo também na forma como este facilita a sua utilização por utilizadores sem conhecimentos específicos . Os actuais sistemas de perguntas e respostas tentam facilitar o seu uso ao máximo permitindo que qualquer utilizador que não tenha qualquer informação sobre o conteúdo da colecção de dados possa usar o sistema e obter informação relevante.

Hoje em dia os sistemas de recuperação de informação têm como objectivo encontrar documentos que contenham informação relevante para um qualquer assunto. A forma como os utilizadores usam os sistemas é através do uso de interrogações ao sistema. Estas interrogações tentam exprimir o assunto ou tópico que o utilizador pretende pesquisar, tornando-se por vezes complicado obter uma interrogação que consiga exprimir exactamente aquilo que o utilizador pretende pesquisar.

Os sistemas de recuperação de informação tradicionais encontram os documentos relevantes para o tema pretendido pelo utilizador, no entanto, o utilizador tem que consultar todo o documento para encontrar a informação pretendida, sendo este um dos grandes problemas dos sistemas de recuperação de informação.

Por vezes é necessário obter informação de uma forma mais rápida e directa. Os sistemas de perguntas e respostas são considerados como sistemas de recuperação

[1] www.google.com

de informação, no entanto têm um comportamento um pouco diferente dos sistemas tradicionais de Recuperação de Informação. Estes sistemas tentam encontrar uma resposta concreta para uma pergunta feita em língua natural pelo utilizador em vez de fornecer um conjunto de documentos que possam ter a resposta à pergunta. Com estes sistemas de perguntas e respostas, o utilizador deixa de procurar informação sobre um tema e passa a procurar uma resposta para uma pergunta concreta feita em língua natural, eliminando assim um dos grandes problemas dos sistemas tradicionais.

A principal motivação para o desenvolvimento e implementação de um sistema de perguntas e respostas para a Língua Portuguesa é responder à necessidade cada vez maior que os utilizadores têm de obter respostas concretas de uma forma simples.

Um outro problema dos sistemas tradicionais de recuperação de informação é a necessidade do uso de interrogações num formato específico, restringindo assim o uso destes sistemas por parte de muitos utilizadores que não têm conhecimentos suficientes para usar estes sistemas. O facto dos sistemas de recuperação de informação apresentarem aos utilizadores um conjunto de documentos que possam ter a informação desejada também restringe o uso dos sistemas por parte de muitos utilizadores, pois nem todos querem perder tempo à procura da informação necessária num conjunto de documentos.

Com o uso de sistemas de perguntas e respostas estes problema tentam ser resolvidos, pois as perguntas são feitas em língua natural e os resultados apresentados concretos, permitindo assim que qualquer utilizador possa usar um sistema destes para fazer uma pergunta em língua natural e obter uma resposta concreta.

A língua Portuguesa é uma das línguas mais faladas em todo o mundo, sendo a oitava língua mais falada. Existem entre 170 e 210 milhões de pessoas a falarem Português em todo o mundo. Apesar de ser uma língua muito usada existem muito poucos sistemas de Perguntas e Respostas feitos para responderem a perguntas em Língua Portuguesa.

Ao serem implementados sistemas de perguntas e respostas para a língua Portuguesa, serão analisados documentos em Português que não seriam analisados por sistemas preparados para outras línguas, permitindo assim que uma grande quantidade de utilizadores possa ter acesso a mais informação de uma forma mais fácil e simples.

1.2 Objectivos

O objectivo deste trabalho é mostrar que é possível implementar um sistema de perguntas e respostas para a língua Portuguesa.

Com este trabalho pretende-se implementar um protótipo dum sistema de perguntas e respostas para a Língua Portuguesa que consiga responder a algumas perguntas simples em língua Portuguesa através da pesquisa da resposta numa colecção de documentos sem restrições de domínio.

Como será detalhado em capítulos posteriores, a implementação deste protótipo tenta usar algumas abordagens diferentes das que são usadas pela maior parte dos sistemas de perguntas e respostas, usando alguns métodos diferentes e inovadores em alguns componentes do sistema

Outro objectivo deste deste é dar inicio a uma área de investigação no Departamento de Informática da Universidade de Évora para que um dia se consiga transformar o protótipo dum sistema de perguntas e respostas numa aplicação concreta que possa ser usada por qualquer utilizador para responder a perguntas feitas em Língua Portuguesa sobre colecções de documentos escritos em Português.

1.3 Contribuição

Nesta secção tento descrever qual a contribuição efectuada efectuada no âmbito deste trabalho para o sistema de perguntas e respostas desenvolvido no Departamento de Informática da Universidade de Évora.

Efectivamente este sistema de perguntas e respostas foi elaborado por um conjunto de pessoas do Departamento de Informática da Universidade de Évora, existindo componentes do sistema desenvolvidos por grupos diferentes.

A minha contribuição neste sistema está espalhada um pouco por todo o sistema, uma vez que a definição da arquitectura global e a integração de todas as partes do sistema de forma a que se tornassem compatíveis ficou a meu cargo. A minha contribuição para este trabalho passou também pelo estudo das várias alterações que os módulos tiveram que sofrer de forma a que estes pudessem ser interligados entre si com o objectivo de formar um sistema de perguntas e respostas.

Em particular, fiz todo o estudo necessário para fazer a integração do sistema de recuperação no resto do sistema de perguntas e respostas bem como a sua integração e toda a parte de pré-processamento das perguntas e dos documentos.

Em termos de desenvolvimento de módulos, aquele para que mais contribuí foi na implementação do módulo de interpretação semântica das perguntas e das respostas bem como a sua integração com o resto, sendo todo este módulo implementado no âmbito deste trabalho.

Este sistema de perguntas e respostas participou numa conferencia de sistemas de recuperação de informação que faz a avaliação dos sistemas participantes. Para que este sistema pudesse participar nessa conferencia foi necessário fazer algumas adaptações, principalmente na forma como o sistema recebia as perguntas e na forma como mostrava as respostas, ficando toda esta adaptação do sistema a meu cargo.

Por fim, foi também feito no âmbito desta dissertação um estudo sobre os resultados obtidos pelo sistema de forma a identificar os seus problemas e encontrar as suas causas para que este pudesse ser melhorado num futuro próximo.

1.4 Estrutura

Esta dissertação tenta descrever o trabalho de investigação feito para o desenvolvimento de um sistema de perguntas e respostas para a Língua Portuguesa. Nesta secção são descritas as várias partes desta dissertação.

No capitulo 1 é feita uma introdução do trabalho feito , a sua importância, os seus objectivos bem como os objectivos pretendidos.

No capitulo 2 é analisada a história da evolução dos sistemas de perguntas e respostas ao logo dos anos, fazendo-se a descrição de alguns sistemas que marcaram a evolução deste tipo de sistemas. Neste capitulo são também descritas algumas conferências onde são analisados sistemas deste tipo, sendo também descritos alguns sistemas de perguntas e respostas que participaram na ultima edição de uma das conferencias.

No capitulo 3 é descrito o sistema de perguntas e respostas implementado, descrevendo-se todos os pormenores do sistema concreto. Neste capítulo são também descritas algumas aplicações reais que o sistema teve.

No capitulo 4 é feita uma avaliação critica do sistema de perguntas e respostas que foi implementado, mostrando-se os pontos fortes, os problemas do sistema e

a resolução destes mesmos problemas. Neste capitulo são também analisados os resultados da avaliação do sistema por parte de uma entidade externa, mostrando assim a validade dos resultados do sistema.

No capitulo 5 é feita uma conclusão sobre todo o trabalho, analisando-se de uma forma muito geral os pontos fortes do sistema, os pontos fracos, a resolução de alguns problemas. Neste capitulo é também feita uma pequena referência a algum trabalho futuro sobre este sistema.

Capítulo 2

Estado da arte

2.1 Introdução

Os sistemas de perguntas e respostas são sistemas de recuperação de informação. Estes sistemas têm como objectivo fornecer respostas precisas a perguntas em língua natural, enquanto que o objectivo dos sistemas de recuperação de informação tradicionais é fornecer aos utilizadores uma lista de documentos relevantes para uma interrogação feita numa linguagem formal muito simplificada.

Estes sistemas têm vindo a ser estudados e desenvolvidos dentro da área do processamento da língua natural desde o ano de 1970. O LUNAR foi um dos primeiros sistemas de perguntas e respostas que apareceu e permitia a geólogos fazerem perguntas sobre amostras de rochas e solos que foram recolhidas na lua durante a missão Apollo 11 [1].

Com o aparecimento de algumas conferências sobre sistemas de recuperação de informação como o TREC[1] e o CLEF[2] houve uma grande motivação pela parte de muitos investigadores que levou à implementação de sistemas de perguntas e respostas em várias línguas. Estas conferências têm como principal objectivo avaliar vários tipos de sistemas de recuperação de informação como os sistemas de perguntas e respostas.

Com o desenvolvimento e investigação das equipas participantes nestas conferencias, houve uma grande melhoria nos resultados obtidos pelos sistemas perguntas e respostas.

Estas conferências são os acontecimentos mais importantes na área da recuperação de informação, sendo os resultados obtidos no TREC e no CLEF tão

[1]Text REtrieval Conference

[2]Cross Language Evaluation Forum

importantes que são considerados como o estado da da área de recuperação de informação.

2.1.1 Os primeiros sistemas de perguntas e respostas

Os primeiros sistemas de Perguntas e Respostas eram sistemas altamente complexos. que transformavam as perguntas em língua natural em interrogações que podiam ser executadas numa base de conhecimento. A transformação das perguntas em língua natural dava resultado a uma interrogação num formato especifico para uma determinada base de conhecimento. Esta nova representação das perguntas era executada na base de conhecimento com o fim de encontrar uma resposta para a pergunta em língua natural.

Estes sistemas de perguntas e respostas eram bastante limitados pois trabalhavam sobre um domínio especifico de informação. As bases de conhecimento com que estes sistemas trabalhavam eram bastante complicadas de elaborar, tornando assim os sistemas muito pouco escaláveis.[2]

Na secção 2.2 são descritos mais em pormenor alguns dos primeiros sistemas de perguntas e respostas mais importantes, como por exemplo o LUNAR e SHRDLU[2].

Desde o inicio da investigação da área de processamento de língua natural, a tarefa de responder a perguntas em língua natural era um dos pontos mais importantes(Greem, Wolf, Chomsky, & Laughry, 1963; Simmons, 1965, 1970)[3], sendo a análise sintáctica e a interpretação semântica os factores mais importantes para se conseguir obter respostas para perguntas feitas em língua natural.

A investigação nesta época tentava resolver problemas complexos relacionados com a interpretação semântica, representação de conhecimento e com a inferência. Os sistemas desenvolvidos na altura conseguiam ter boas interpretações e inferências semânticas, no entanto era muito difícil fazer a adaptação destes sistemas a novos textos ou a novos domínios, pois os sistemas estavam desenhados para trabalharem sobre um determinado domínio e muitas das vezes estavam optimizados para reponderem a perguntas sobre esse domínio.

2.1.2 Novas abordagens

Recentemente, com o aparecimento de novas técnicas de processamento de informação, houve um grande desenvolvimento dos sistemas de perguntas e respostas em domínios abertos.

Com esta grande evolução, os sistemas de perguntas e respostas começaram a ser usados em qualquer contexto e a usar novas bases de conhecimento.

Uma das grandes diferenças entre estes sistemas de perguntas e respostas e os primeiros é a facilidade com que podem ser alterados para que o sistema possa trabalhar com novos domínios.

Foi principalmente com o aparecimento de uma tarefa de sistemas de perguntas e respostas no TREC [3], organizada pelo NIST[4], que a pesquisa e desenvolvimento na área dos sistemas de Perguntas e Respostas teve um grande crescimento e começaram a obter resultados bastante interessantes.

Na tarefa de sistemas de perguntas e respostas do TREC, a base de conhecimento é substituída por uma grande colecção de documentos de texto retirados de jornais e revistas. Com a substituição da base de conhecimento por documentos de texto, é eliminado um dos grandes problemas dos primeiros sistemas de perguntas e respostas.

Esta tarefa apareceu na edição 8 do TREC em 1999 (TREC-8), e o seu principal objectivo é avaliar e comparar sistemas que consigam responder com uma resposta exacta a perguntas feitas em língua natural num domínio não especifico. Para isso, a pesquisa das resposta deve ser feita numa colecção de documentos de texto reais que são retirados de jornais e revistas.

As primeiras edições do TREC fomentaram bastante o desenvolvimento e investigação de sistemas de Perguntas e Respostas, no entanto este desenvolvimento verificou-se mais para a língua Inglesa do que para outras línguas. Para outras línguas houve muito pouca investigação, reduzindo muito a qualidade destes sistemas.

As últimas edições do TREC têm sido cada vez mais exigentes em relação ás perguntas. Este aumentar de exigência foi bom para os sistemas de perguntas e respostas, fazendo com que os sistemas participantes sofressem uma grande evolução para satisfazer os novos graus de exigência. Com este aumentar de exigência ao longo dos anos os sistemas de perguntas e respostas tornaram-se cada vez melhores e mais fiáveis.

Estes sistemas usam várias técnicas diferentes para conseguirem responder às perguntas, no entanto, a maior parte usa em conjunto técnicas de recuperação de informação clássicas junto com algumas técnicas de processamento de língua natural[2].

[3]Text Retrieval Conference

[4]National Institute of Standards and Technology

Com o uso destas técnicas de processamento de língua natural e o uso de vários processos de recuperação de informação, os sistemas de perguntas e respostas começaram a ter mais sucesso e um maior valor de confiança nas respostas.[2]

2.2 Evolução dos sistemas de Perguntas e Respostas

Os primeiros sistemas de perguntas e respostas que apareceram foram o WOLFIE, o LUNAR, o SHRLDU.

O WOLFIE é um sistema que consegue "aprender" um léxico ou um sistema de processamento de língua natural que consegue transformar interrogações a bases de dados numa forma lógica executável. Este sistema recebe apenas como dados de entrada um conjunto de perguntas previamente anotadas.

O LUNAR deve ter sido o primeiro sistema de perguntas e respostas real. Este sistema foi desenhado para conseguir responder a perguntas sobre amostras de rochas e solos que foram recolhidos durante a missão Apollo 11. Este sistema é bastante limitado devido à sua base de conhecimento, conseguindo trabalhar apenas neste domínio.

Este sistema de perguntas e respostas respondia às perguntas fazendo uma transformação da pergunta em língua natural para uma interrogação de bases de dados. Esta interrogação era depois executada na base de dados com o fim de obter uma resposta.

O SHDRLU é um sistema que faz a simulação de um robot que consegue manipular objectos num mundo virtual. Este sistema consegue também responder a algumas perguntas sobre as posições dos blocos no mundo virtual e justificar a sua posição. Para fazer a manipulação dos objectos, o sistema recebe do utilizador ordens em língua natural.

O SHRDLU pode ser considerado como um sistema de processamento e compreensão de língua natural.

Estes foram alguns dos sistemas mais importantes na evolução dos sistemas de perguntas e respostas. Nas próximas secções estes sistemas são descritos mais em detalhe, sendo exploradas algumas das suas características bem como as tecnologias e teorias que usam.

2.2.1 WOLFIE - O primeiro sistema de Perguntas e Respostas

O WOLFIE[5][4] foi um dos primeiros sistema de perguntas e respostas. Este sistema tem como principal característica conseguir adquirir informação semântica a partir das palavras e dos seus significados.

O objectivo deste sistema é a construção automática de um modelo léxico para um sistema integrado de PLN que consegue adquirir informação léxica e gerar analisadores sintácticos para serem usados como interfaces de outras aplicações que trabalham com a língua natural.

O resultado produzido pelo WOLFIE pode ser usado para ajudar outros sistemas, como por exemplo o CHILL[6][Zelle, 1995][3].

Os dados que o CHILL recebe é um conjunto de frases associadas à sua representação semântica. O analisador sintáctico que é "aprendido" pelo CHILL a partir do conjunto de dados é capaz de transformar as frases usadas para o treino na sua representação correcta, bem como fazer uma generalização para interpretar e fazer a representação de novas frases.

Os analisadores sintácticos produzidos pelo CHILL, convertem as perguntas em língua natural para interrogações em Prolog. Estas interrogações vão depois ser analisadas para se obter uma resposta para a pergunta(Zekke &Mooney 1996).

De seguida, são apresentadas algumas perguntas em língua natural junto com a sua representação em Prolog:

```
What is the capital of the state with the biggest population?
answer(C, (capital(S,C), largest(P,(state(S),population(S,P))))).

What state is Texarkana located in?
answer(S, (state(S), eq(C,cityid(texarkana,_)), loc(C,S))).
```

Quando é fornecido ao CHILL um conjunto de frases em língua natural e a sua representação em Prolog, o CHILL é capaz de *aprender* um analisador sintácticos que consegue fazer a análise de novas perguntas em língua natural e gerar uma interrogação de base dados que representa a frase em língua natural. Usando o WOLFIE, a informação léxica pode ser fornecida automaticamente ao CHILL e criada uma ligação entre as palavras, os predicados e termos das interrogações.

[5]WOrd Learning From Interpreted Examples

[6]Consructive Heuristics Induction for Language Learning

Um sistema capaz de aprender uma forma léxico-semântica apresenta um conjunto de frases. Cada uma destas frases tem uma lista de palavras ligadas a uma árvore com uma representação semântica. O objectivo destes sistemas é obter uma interpretação léxico-semântica que seja consistente com os dados.

Esta representação léxico-semântica consiste num par (oração, significado), por exemplo `([biggest], largest( _,_))`, onde as orações e o seu significado são retirados das frases e da sua representação fornecida ao sistema.

A representação de cada frase pode ser composta por vários componentes, sendo cada um escolhido a partir das possíveis representações das orações que aparecem várias vezes na mesma frase.

Um dos objectivos do WOLFIE é minimizar a ambiguidade e o tamanho do léxico "aprendido", podendo desta forma ser melhorada a precisão e facilitar o processo de geração do analisador sintáctico.

Este sistema permite que a mesma frase tenha vários significados(homónimos), que várias frases tenham o mesmo significado(sinónimos), e que algumas orações duma frase não tenham qualquer significado.

Este sistema assume que o significado de uma frase é composto pelo conjunto dos significados das várias orações que compõem a frases.

Neste sistema é também assumido que cada parte individual das frases se deve apenas a uma palavra ou a uma oração da frase, e que o significado de uma palavra aparece pelo menos uma vez na representação de uma frase. A representação de uma oração neste sistema é um sub-grafo da representação da frase, que por sua vez também é um grafo.

2.2.2 LUNAR

O LUNAR[5][6] foi um dos primeiros sistemas de perguntas e resposta que apareceu. Este sistema de perguntas e respostas foi usado para fazer perguntas sobre amostras de rochas e solos lunares.

Neste sistema, a informação estava armazenada numa base de dados que tinha toda a informação sobre as amostras das rochas e solos.

No LUNAR, as perguntas eram feitas em língua natural Inglesa e depois eram traduzidas para uma interrogação de base de dados que pudesse ser executada na base de dados que tinha toda a informação das rochas e dos solos. Para fazer esta tradução foram usadas algumas regras sintácticas e semânticas.

Depois da pergunta em língua natural estar traduzida numa interrogação de base de dados, esta era executada na base de dados para se tentar obter uma

resposta.

Este sistema tem um analisador sintáctico para um sub-conjunto da língua Inglesa de tamanho desconhecido, consegue fazer a resolução de algumas anáforas, e tem um vocabulário com cerca de 3500 palavras.

ATN's Augmented Transition Networks

O sistema LUNAR é um interface para uma base de dados que usa *ATN's*[7] e a semântica procedimental de Woods para fazer a tradução das perguntas em língua natural numa linguagem que possa ser executada na base de dados. As *ATN's* são uma peça de software que é capaz de usar gramáticas bastante poderosas para processar a sintaxe[7].

Este sistema permite especificar o conhecimento sobre o domínio da aplicação sob a forma de *Extended Transition Networks.*

Para além das *transition networks*, existe uma forma para especificar a forma como a pesquisa vai ser feita nas redes de transição com o fim de obter uma solução para os problemas.

Este sistema foi buscar o seu nome à base de dados que estava a usar e que continha informação sobre amostras de rochas e solos lunares que foram recolhidas durante a missão Apollo 11.

O LUNAR foi demonstrado pela primeira vez na *Second Annual Lunar Science Conference* em 1971. Este sistema tem um desempenho bastante impressionante, sendo capaz de responder correctamente a cerca de 78% das perguntas que lhe são feitas.

Este sistema foi usado por utilizadores que exigiam uma grande precisão. A pesquisa e a investigação de sistemas parecidos com o LUNAR foi continuada por mais de uma década, no entanto os sistemas obtidos eram todos eles muito parecidos com o LUNAR e os resultados obtidos eram muito equivalentes aos resultados do LUNAR.

2.2.3 SHRDLU

O SHRDLU é um programa de computador de processamento e compreensão de língua natural desenvolvido por Terry Winograd no MIT [8].

[7] Augmented Transition Networks

[8] Massachusetts Institute of Technology

Este sistema permite uma interacção com os utilizadores usando a língua natural Inglesa.

O sistema simula um robot que é capaz de manipular um conjunto de objectos numa mesa. Os utilizadores ordenam ao sistema para mover vários objectos num pequeno "mundo de blocos" que contem vários blocos, cones e bolas.

Este sistema tornou-se único devido a quatro simples ideias que fizeram com que a simulação da compreensão fosse bastante mais convincente. Uma destas ideias foi usar um dicionário de palavras de tamanho reduzido. Com o uso deste dicionário de pequenas dimensões torna-se possível interagir com o sistema para dar qualquer ordem ao robot com um conjunto de apenas 50 palavras. O uso dum dicionário de pequenas dimensões facilita bastante o processo de "compreensão" das ordens dadas ao sistema.

Nomes como "bloco", "cone", verbos como "place on" ou "move to", adjectivos como "grande" e "azul" chegavam para fazer a descrição, a localização e a movimentação dos objectos.

A combinação destas palavras formavam uma linguagem básica que o programa conseguia interpretar bem. A interpretação que o sistema faz às ordens do utilizador está quase sempre de acordo com ordem que o o utilizador realmente pretende.

O facto de o SHRDLU ter um vocabulário bastante reduzido é uma das principais razões para o bom desempenho do sistema.

Outra das características que tornaram o SHRDLU único é a sua memória básica onde é possível fornecer um contexto ao sistema.

Se for dada a ordem "põe o cone verde no bloco vermelho" e de seguida, "tira o cone", o sistema considera que o cone é o cone verde que se tinha "falado" anteriormente. O SHRDLU tem a capacidade de procurar na memória básica até encontrar o contexto correcto que possa ser usado para satisfazer uma ordem.

O sistema permite também fazer perguntas sobre a história do discurso, podendo-se perguntar ao sistema perguntas do tipo:

- "agarraste alguma coisa antes do cone?".

Um efeito directo desta memória é que o sistema pode responder a perguntas sobre o que é ou não possível de fazer no seu pequeno mundo. O SHRDLU pode por exemplo deduzir que os blocos podem estar empilhados olhando para as ordens que já foram dadas ao sistema, mas consegue também perceber que não pode haver blocos empilhados em cima de pirâmides depois de já ter havido uma ordem para empilhar um bloco em cima de um cone e ter verificado que o bloco caiu.

O pequeno mundo do SHRDLU contem algumas regras básicas de física que fazem com que alguns blocos quando empilhados em cima de outros caiam sob determinadas situações.

A última característica que fez deste sistema único é a capacidade de se lembrar dos nomes que foram dados aos objectos.

Extracto de uma *Conversa* com SHRDLU

De seguida está um excerto da famosa demonstração do SHRDLU:

- Person: PICK UP A BIG RED BLOCK.
- Computer: OK.
- Person: GRASP THE PYRAMID.
- Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
- Person (changing their mind): FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
- Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
- Computer:OK.
- Person:WHAT DOES THE BOX CONTAIN?
- Computer:THE BLUE PYRAMID AND THE BLUE BLOCK.
- Person:WHAT IS THE PYRAMID SUPPORTED BY?
- Computer:THE BOX.
- Person:HOW MANY BLOCKS ARE NOT IN THE BOX?
- Computer:FOUR OF THEM.
- Person:: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
- Computer: YES, THE RED CUBE. ....

O SHRDLU é composto principalmente por 4 elementos básicos: um analisador sintáctico, uma gramática de Inglês, programas para análise semântica que servem para transformar uma frase em língua natural numa sequência de comandos que possam ser executadas pelo robot, e um subsistema que sabe arrumar os vários blocos.

A gramática usado pelo SHRLDU é baseada na noção das gramáticas sistémicas. As gramáticas sistémicas são um conjunto de redes que especificam as características de cada uma das unidades sintácticas, a forma como cada uma das unidades funciona e a forma como as várias unidades estão interligadas.
Este sistema ao fazer uma integração do conhecimento da sintaxe e da semântica consegue evitar algumas escolhas que possam vir a ter ambiguidades. O analisador sintáctico usado pelo sistema dá preferência a escolher uma das primeiras escolhas do que fazer "back-trackings" exaustivos.

O SHRDLU teve bastante sucesso na área da Inteligência Artificial, levando a muitos investigadores a terem um optimismo excessivo que depressa se perdeu quando os sistemas começaram a lidar com situações mais reais num mundo real, complexo e cheio de ambiguidades.

2.3 Sistemas recentes

Os sistemas de perguntas e respostas actuais podem ser divididos em várias classes. Esta separação dos sistemas de perguntas e respostas pode ser feita através das várias características que os sistemas apresentam.

Actualmente existem muitos tipos de sistemas de perguntas e respostas, podendo ser categorizador por várias formas: 1)pela língua natural para o qual foram implementados e pelo número de línguas que pode usar, 2) pelo tipo de base de conhecimento que é usada para pesquisar as respostas e 3) pelo domínio do sistema de perguntas e respostas.

1) A maior parte dos sistemas de perguntas e respostas estão feitos para trabalharem apenas com uma língua, chamados mono-lingues. No entanto, os sistemas de perguntas e respostas podem ser bi-lingues, ou multi-lingues. Os sistemas de perguntas e respostas bi-lingues respondem às perguntas numa língua natural com respostas numa outra língua natural, enquanto que os sistema multi-lingues respondem a uma pergunta numa língua natural com respostas em várias línguas naturais, sendo normalmente dada uma resposta na língua natural da pergunta.

Existem também outros sistemas de perguntas e respostas que são independentes da língua. Estes sistemas não têm qualquer restrição em relação à língua da pergunta como à língua da base de conhecimento. Estes sistemas podem ser muito flexíveis, mas, no entanto existem muito poucos sistemas deste tipo devido à sua grande complexidade.

2) Outra forma de fazer uma separação dos sistemas de perguntas e respostas é pela representação da base de conhecimento do sistema. A base de conhecimento dos sistemas de perguntas e respostas normalmente está representada por uma enorme colecção de documentos de textos, no entanto a base de conhecimento pode também estar representada numa base de dados. Outra forma de base de conhecimento que ultimamente tem vindo a ser a usada é a própria WEB, chegando mesmo a ser considerada a maior base de conhecimento existente no mundo.

3) O domínio dos sistemas de perguntas e respostas é também muito importante. Os sistemas de perguntas e respostas trabalhavam sobre um domínio especifico, respondendo apenas a perguntas dum certo domínio. Estes sistemas eram bastante limitados pois era muito complicado alterar o sistema para que este consegui-se trabalhar sobre outros domínios.

   Hoje em dia, com o aparecimento da "track QA" do "TREC", os sistemas de perguntas e respostas deixaram de trabalhar sobre domínios específicos para trabalharem sobre domínios abertos, podendo assim responder a perguntas sobre um domínio qualquer desde que a pergunta pudesse ser respondida com informação disponível na base de conhecimento.

Embora todos estes tipos de sistemas de perguntas e respostas tenham características bastante diferentes, normalmente usam uma arquitectura global bastante parecida dentro do mesmo tipo de perguntas e respostas. Mesmo em sistemas de perguntas que possam ser considerados como sendo de tipos diferentes, pode haver uma arquitectura global do sistema muito parecida.

Normalmente os sistemas de perguntas e respostas são compostos por vários módulos que interagem entre si. Estes módulos, na maior parte dos sistemas de perguntas e respostas têm o mesmo objectivo, no entanto a forma como são implementados, as tecnologias que usam e as teorias usadas na implementação destes módulos podem ser completamente diferentes.

Uma das fases mais importantes nos sistemas de perguntas e respostas é a análise e o processamento das perguntas em Língua Natural, que, normalmente faz uma análise à pergunta usando uma grande variedade de ferramentas, para depois gerar uma nova representação da pergunta que é mais fácil de interpretar pelo sistema.

Um dos módulos que normalmente todos os sistemas de perguntas têm é um módulo que faz a pesquisa de informação na base de conhecimento ou na colecção de documentos que possa ser relevante para a pergunta e que possa ter informação para gerar uma resposta. Este módulo de pesquisa de informação relevante para a pergunta pode ser feito de várias formas usando várias técnicas, mas, normalmente é feito com o uso de técnicas de normais de recuperação de informação. A forma como este módulo é implementado pode variar bastante com a forma como da base de conhecimento está representada, sendo normalmente usados sistemas de recuperação de informação já existentes.

Por fim, um outro módulo que todos os sistemas de perguntas e respostas têm, independentemente do seu tipo, é um módulo de geração das respostas. Este módulo normalmente recebe a nova representação da pergunta, a interpretação da pergunta e a informação que possa ser relevante para a pergunta.

Com estes dados, o módulo de geração das respostas vai tentar gerar uma resposta para pergunta com o uso de varias pesquisas sobre a informação relevante para a pergunta através de várias técnicas. As técnicas que são usadas variam muito com o tipo de representação que foi feita para as perguntas. Normalmente estas técnicas de geração das perguntas usam sistemas que fazem algum tipo de "pattern matching", técnicas de processamento de língua natural ou ambas as técnicas.

As técnicas usadas na geração das respostas podem variar muito com o tipo da representação escolhida para as perguntas, no entanto, pode haver metodologias completamente diferentes em sistemas que usem uma representação idêntica para as perguntas.

Os sistemas de perguntas e respostas, ao fazerem a análise das perguntas criam uma representação mais simplificada das perguntas para que esta possa ser "compreendida" pelo sistema. Esta representação normalmente é uma representação quase canónica da pergunta onde a pergunta reduzida a uma forma muito mais simples. Esta transformação das perguntas simplifica bastante a representação da pergunta, no entanto, esta representação tem de conseguir manter o mesmo significado da pergunta original.

Para os sistemas de perguntas e respostas "perceberem" as perguntas é muitas vezes calculado o tipo da pergunta bem como o tipo da resposta. Para se calcular o tipo da pergunta e da resposta são usadas várias técnicas.

Uma das técnicas para se descobrir qual o tipo da pergunta bem como o tipo da pergunta é descobrir qual o foco da pergunta. Ao conseguir-se descobrir qual o foco da pergunta ou a palavra chave da pergunta consegue-se obter muita informação sobre a pergunta.

Para se encontrar o foco da pergunta ou mesmo o tipo da pergunta ou o tipo da resposta que deve ser dada à pergunta são usadas classes gerais de perguntas.

Estas classes gerais de perguntas tentam representar duma forma geral muitas das perguntas em Língua Natural que são feitas no dia a dia. A construção destas classes é feita manualmente através da consulta e análise a perguntas reais e tenta agrupar as perguntas que são feitas mais frequentemente em várias classes, de forma a que uma única classe de perguntas consiga representar muitas perguntas que tenham um mesmo formato ou significado.

Com a criação de classes que representam grande parte das perguntas, é então possível descobrir qual o tipo que as perguntas têm, bem como o tipo da resposta que deve ser dada. Através de regras de "pattern matching" que fazem uma análise a todas as classes de perguntas existentes e às perguntas em língua natural consegues-se obter o tipo da pergunta e da resposta. Para se conseguirem estes tipos é feita uma análise a várias partes da classe da perguntas que possa representar a pergunta e uma análise a várias partes da pergunta em língua natural. Esta análise tenta "unificar" algumas palavras chave nas classes das perguntas com algumas palavras chave na pergunta em Língua Natural. Quando este "match" é feito é encontrada uma classe que representa um tipo de perguntas, e, que em especifico representa a pergunta em questão. O sistema ao encontrar a classe que representa a pergunta consegue identificar tipo da pergunta e o tipo da resposta, pois associado a cada classe de perguntas está um tipo que será o tipo de todas as perguntas que pertencem à classe e um tipo que será o tipo das respostas que devem ser dadas às perguntas que pertencem a essa classe.

Outra possível forma para os sistemas conseguirem perceber as perguntas feitas em Língua Natural é o uso de técnicas de processamento de língua natural para criar estruturas semânticas que possam representar as perguntas como as redes lexicais ou as DRSs.

Normalmente na análise e processamento das perguntas são usadas várias ferramentas linguísticas. Uma das ferramentas linguísticas mais usadas e mais importantes nesta parte dos sistemas de perguntas e respostas são os analisadores sintácticos que identificam o tipo de cada uma das palavras das perguntas e retiram toda a informação de cada uma das palavras das perguntas em Língua Natural. Estes analisadores sintácticos são a ferramenta linguística mais importante no processamento das perguntas.

Uma das fases que também é bastante importante, é a recuperação de informação relevante para a pergunta. Esta fase não é essencial em todos os sistemas de

perguntas e respostas mas facilita bastante a procura das respostas quando a colecção de documentos é muito grande.

A implementação do módulo de recuperação de informação varia muito com a forma como a base de conhecimento está representada. Normalmente a base de conhecimento disponível pelo sistema de perguntas e respostas é composta por uma grande colecção de documentos de texto. Neste caso são utilizadas ferramentas de recuperação de informação que fazem a pesquisa de documentos de texto da colecção que possam ser relevantes para a pergunta. Normalmente as ferramentas de recuperação de informação usadas para fazer a recuperação de informação são aplicações de recuperação de informação já existentes.

Os sistemas de recuperação de informação, para fazerem as pesquisas precisam sempre de uma interrogação para fazer a pesquisa da informação nos documentos. Estas interrogações normalmente são construídas a partir de algumas palavras chave que são retiradas da pergunta original em língua natural ou da representação simplificada da pergunta que foi feita no módulo de processamento das perguntas. Com a execução destas interrogações nos sistemas de perguntas e respostas consegue-se então obter a informação relevante para as perguntas.

Uma outra forma de representar a base de conhecimento são as bases de dados. Quando a base de conhecimento do sistema está representada numa base de dados, são feitas interrogações às bases de dados para se obter a informação relevante para a pergunta. Neste caso, as interrogações que vão ser feitas à base de dados são construídas da mesma forma que no caso anterior onde a base de conhecimento está representada por um conjunto de ficheiros.

Uma outra forma de base de conhecimento que é bastante usada em sistemas de perguntas e respostas é a WEB. Neste caso a base de conhecimento é composta por todos os documentos disponíveis na WEB. Para se fazer a recuperação de informação que possa ser relevante para a pergunta, são normalmente usados motores de busca de informação sobre a WEB como o Google ou o Altavista.

Estes motores de busca sobre a WEB, tal como os motores de recuperação de informação sobre colecções de ficheiros, ou os sistemas de gestão de bases de dados precisam de interrogações para fazer a busca da informação pretendida. Para isso são criadas as interrogações com base na pergunta para serem executadas no motor de busca.

Muitos sistemas de perguntas e respostas fazem apenas a recuperação dos documentos que são relevantes para a pergunta, sendo toda a pesquisa das respostas feita sobre todo o documento recuperado, no entanto, existem alguns sistemas, que, depois de obterem os documentos relevantes para a pergunta fazem a recuperação

de blocos de texto relevantes dentro do documento texto (passage retrieval). Com a obtenção dos blocos de texto relevantes torna-se mais simples a geração das respostas pois apenas vão ser analisados pequenos blocos de texto em vez de todo o documento de texto completo.

Os sistemas de recuperação de informação usados nos sistemas de perguntas e respostas usam normalmente técnicas de pesquisa que tiram partido de sinónimos das palavras e as suas variantes inflecionais e morfológicas. O uso de sinónimos, variantes inflecionais e morfológicas pode ser feito pelo próprio motor de busca ou na geração das interrogações que vão ser feitas ao motor de busca.

Neste ultimo caso, em vez de ser passada uma interrogação com as palavras tal como estão na pergunta em Língua Natural são usadas as várias variantes das palavras originais.

O uso destas variantes na recuperação de informação relevante traz algumas vantagens mas também algumas desvantagens. Com estas variações consegue-se obter mais informação relevante para a pergunta pois é feita uma pesquisa mais geral. Embora se consiga obter mais informação relevante com o uso destas técnicas, vai-se obter informação que pode não ser muito relevante para a pergunta, ou, pelo menos o conjunto total de informação relevante recuperada não é tão relevante como se não fossem usadas as várias variações das palavras. Quando são usadas variantes das palavras deve-se encontrar um meio termo para que se consiga encontrar alguma informação relevante que seja bastante relevante, e não, encontrar muita informação que seja pouco relevante.

O objectivo final dos sistemas de perguntas e respostas é dar uma resposta para uma pergunta feita em língua natural.

Este processo pode ser feito de várias formas dependendo da forma como as perguntas foram tratadas e a da forma como a base de conhecimento está representada.

Embora existam grandes diferenças na forma como as respostas são extraídas da base de conhecimento, a forma como é retirada pode ser muito parecida, pois, na maior parte dos sistemas de perguntas e respostas, a resposta é extraída através de uma *unificação* entre a representação da pergunta e a informação relevante para a pergunta.

Esta *unificação* pode ser feita de formas muito diferentes, pois está dependente da representação interna das perguntas perguntas e da forma como a base de conhecimento está representada.

A unificação entre a pergunta e uma possível resposta pode ser baseada em várias informações da pergunta que possam estar disponíveis, como o tipo da

pergunta, o tipo da resposta que deve ser dada bem como informações especificas de cada uma das palavras da pergunta, como o o seu tipo, género e o número.

Quando o sistema sabe qual é o tipo da resposta que deve ser dada, ou qual o foco da pergunta, é procurado na informação relevante um bloco de texto que tenha o mesmo tipo da resposta, ou que tenha o mesmo foco que a pergunta. Para fazer esta identificação são normalmente usadas as várias classes com os vários tipos de perguntas que foram criadas anteriormente e que conseguem identificar o tipo ou o foco de um bloco de texto.

Existem outros sistemas que usam apenas uma pesquisa superficial nos blocos de informação relevantes. Neste caso são escolhidos blocos de texto que tenham o maior número de palavras relevantes da pergunta o mais perto umas das outras.

De seguida são descritos alguns sistemas de perguntas e respostas de vários tipos que usam várias metodologias diferentes para responder a diferentes tipos de perguntas. Estes sistemas de perguntas e respostas são sistemas recentes que obtiveram bons resultados a responder a perguntas em língua natural.

2.3.1 The Southern Methodist University aproach

Um dos sistemas que participou no "TREC-Q&A" foi o sistema feito por Harabagiu, Pasca e Maiorano, do Department of Computer Science and Engineering da Southern Methodist University[8].

Este grupo tentou fazer uma integração de técnicas de PLN sobre bases de conhecimento com outras técnicas de processamento para resolver dois dos maiores problemas dos sistemas de perguntas e respostas:

- A captura da semântica de perguntas feitas sobre domínios abertos.
- A justificação das respostas.

Este sistema tenta ser bastante preciso nas respostas, facilmente escalável de forma a ser capaz de responder a perguntas mais complexas do que as feitas no TREC.

A qualidade dos sistemas de perguntas e respostas sobre domínios abertos pode ser bastante melhorada pelo uso de várias técnicas de PLN. As técnicas de processamento necessárias pelos sistemas de perguntas e respostas devem ser bastante distintas das técnicas que são usadas nos sistemas de recuperação de informação que simplesmente obtêm um conjunto de documentos com a informação necessária.

Este sistema, para conseguir trabalhar sobre um domínio aberto sistema substitui o "pattern-matching" dos sistemas de recuperação de informação, por métodos que dependem do reconhecimento do tipo da pergunta e do tipo que deve ser a resposta.

Classificação das perguntas

A classificação das perguntas é feita através do reconhecimento de algumas palavras chaves da pergunta(que, quanto, quem, ...). O processamento das perguntas, inclui também a identificação de algumas palavras essenciais para a pergunta, usando para isso alguns métodos empíricos baseados num conjunto de heurísticas que operam sobre o o resultado da análise da pergunta. As palavras identificadas por esta fase como sendo importantes para a pergunta são enviadas para o motor de busca para obter a informação relevante para a pergunta.

A precisão de todo o sistema depende bastante do foco da pergunta, pois a extracção da pergunta é baseada no foco da pergunta. O foco da pergunta é muito importante para este sistema, pois a resposta para a pergunta vai ser centrada numa palavra de um documento que se identifique com o foco da pergunta.

Recuperação de informação

Para que a extracção da pergunta seja mais rápida, o motor de busca devolve apenas os parágrafos que contêm todas as palavras relevantes da pergunta. Estes parágrafos são ordenados de forma a que as palavras relevantes da pergunta estejam o mais próximo possível umas das outras. As respostas são extraídas quando o tipo da pergunta e o tipo de resposta são reconhecidos num parágrafo, sendo depois atribuído uma pontuação às respostas. A atribuição da pontuação das respostas é feita através de várias heurísticas baseadas em análises superficiais aos textos.

Todo este processamento está limitado pelo resultado da anotação dos textos("named entity recognition"), pela classificação semântica do tipo das perguntas baseada na informação fornecida pelo WordNet(Fellbaum 1998) e pelo "parse" de cada uma das frases.

Processamento das perguntas

O módulo de processamento das perguntas identifica qual a classe da pergunta. A identificação da classe da pergunta é feita através do uso dum conjunto de taxonomias de perguntas tipo. Para que o sistema seja capaz de identificar a classe da pergunta através das taxonomias das várias perguntas, é feita uma representação semântica da pergunta que obtêm todas as relações entre as várias orações da pergunta. O reconhecimento da classe da pergunta é feita através da comparação da

representação semântica da pergunta com a representação semântica dos nós da taxonomia das perguntas contidas no conjunto de perguntas padrão. Os nós da taxonomia das perguntas contêm a classe da pergunta o tipo da resposta, o foco da pergunta e a classe semântica de cada palavra da pergunta. Através dos nós da taxonomia da pergunta são geradas algumas combinações palavras chave. Estas combinações de palavras chave são depois usadas como uma interrogação que vai ser executada no motor de busca para obter um conjunto de documentos relevantes para a pergunta, podendo-se assim obter vários conjuntos de respostas que serão ordenados segundo algumas heurísticas descritas em (Moklovan et al.1999).

Com a obtenção de vários conjuntos de respostas para uma pergunta, é aumentada a probabilidade do sistema obter uma resposta correcta para a pergunta, no entanto, onde este sistema ganhou muita qualidade foi no módulo que faz a justificação das respostas que foram obtidas. Todas as respostas obtidas pelo sistema são transformadas numa representação semântica.

Depois de obtida uma representação semântica para as respostas, é feita uma transformação da representação semântica das perguntas e das respostas para uma forma lógica. Esta forma lógica vai então ser passada a um "demonstrador de teoremas" simplificado que vai tentar justificar a resposta. Esta prova tenta criar um "caminho" entre a pergunta e a resposta, sendo depois caminho como justificação da resposta.

No caso do sistema não conseguir justificar uma resposta então essa resposta será eliminada. Com a eliminação destas respostas é eliminado o problema de quando a resposta encontrada não é a resposta que o sistema considera a melhor.

Os resultados que este sistema conseguiu foram bastante bons, conseguindo 77.7% de respostas correctas quando foi usada apenas uma pesquisa superficial, 83,2% de respostas correctas quando foi usada uma base de conhecimento no processamento das perguntas, 77,7% quando foi usada a pesquisa superficial junto com o sistema de justificação das respostas e 89,5% quando foram usados os três sistemas.

2.3.2 CSAIL-MIT

O grupo de investigadores do MIT que implementou este sistema, acredita que para se obter um sistema de perguntas e respostas que consiga responder a vários tipos de perguntas com uma elevada precisão, é necessário uma combinação de várias tecnologias e estratégias.

O sistema CSAIL[9] faz uma abordagem aos sistemas de perguntas e respostas algo diferente, tentando fazer uma integração de técnicas de pesquisa na WEB com técnicas tradicionais de recuperação de informação e com técnicas de anotação de

texto. A integração destas tecnologias num sistema de perguntas e respostas tem o objectivo de aumentar a qualidade das respostas obtidas pelo sistema.

Este tipo de abordagem aos sistemas de Perguntas e Respostas é chamadas de *"knowledge mining"*(Lin and Katz, 2003).

Este sistema foi desenvolvido para participar no TREC, onde as respostas dadas pelos sistemas têm que ser justificadas através da indicação do documento que foi usado para extrair a resposta. Este é um problema para este tipo de sistemas de perguntas e respostas, pois, quando a resposta para uma pergunta é encontrada na Web é necessário encontrar um documento que suporte essa pergunta.

Esta abordagem assume que cada pergunta tem apenas uma resposta, sendo todos os algoritmos desenhados e optimizados para responder com uma grande precisão às perguntas.

Para que o sistema também consiga obter bons perguntas do tipo lista(perguntas cuja resposta é uma lista de itens) foi necessário fazer a alteração dos algoritmos de forma a conseguirem obter resultados equilibrados para os vários tipos de perguntas.

No processamento das perguntas do tipo lista foi feita uma integração de técnicas de pesquisa na Web com técnicas de recuperação de informação e com a anotação de textos.

Este sistema, para responder a perguntas do tipo lista usa um conjunto de módulos interligados entre si. Os módulos existentes neste sistema são:

- recuperação de informação
- recuperação de blocos relevantes(*passage retrievel*)
- extracção das respostas e de remoção de respostas duplicadas.

A ideia deste sistema é eliminar sucessivamente o tamanho da colecção de documentos de onde é possível extrair uma resposta. Este sistema começa por seleccionar uma lista de documentos relevantes para a pergunta e depois escolhe alguns blocos de texto que possam ter informação relevante para a geração da resposta.

No fase de recuperação de informação, o sistema cria uma lista de documentos candidatos que podem ter informação relevante para a geração da resposta. Esta lista de documentos vai depois ser utilizada como entrada dos seguintes módulos do sistema.

Esta é uma das fases mais importantes do sistema, pois se não forem recuperados documentos que sejam relevantes para as perguntas, os módulos das seguintes

fases não vão conseguir obter bons resultados. O sistema que este grupo usou para fazer a recuperação dos documentos foi o Lucene, que é um motor de recuperação de informação livre.

O próximo passo deste sistema de perguntas e respostas é a obtenção de blocos de texto relevantes. A ideia deste processo é reduzir o conjunto de documentos relevantes encontrado pelo sistema de recuperação de informação, obtendo assim um conjunto de blocos de texto que possam ter informação para responder à pergunta.

Para se determinar os blocos de texto para cada um dos documentos relevantes, os documentos são divididos em frases, sendo depois cada uma destas frases pontuadas com base num algoritmo da IBM. Este algoritmo mede a distancia entre as palavras dentro de cada um dos blocos, e soma o número de vezes que as palavras da pergunta aparecem nas frase e o número de palavras que têm o mesmo significado, sendo este obtido pelo WordNet.

O primeiro passo do processo da extracção das resposta é encontrar o foco da pergunta. O foco da pergunta é a palavra ou a frase que é usada para identificar uma classe na ontologia que se identifica com o que se esta à procura na pergunta.

Para ajudar na identificação do foco da pergunta, é criada uma grande base de conhecimento de entidades que contém listas que correspondem aos vários tipos de perguntas existentes. Para a construção esta base de conhecimento, foi recolhida informação sobre os estados dos Estados Unidos, das maiores cidades dos Estados Unidos, de nomes de pessoas, de nomes de países entre outras listas.

Se o tipo da pergunta estiver nestas listas, então o sistema extrai instâncias do tipo da pergunta a partir dos blocos de texto mais relevantes. No caso do tipo da pergunta não estar na base de conhecimento, então são retirados os sintagmas nominais que têm o mesmo tipo que a pergunta ou têm o foco da pergunta como sendo a resposta para a pergunta. No caso de não haver nenhum sintagma nominal com o mesmo tipo que o foco da pergunta, é retirado o sintagma nominal da frase que está mais perto do foco da pergunta.

Normalmente, este sistema extrai respostas duplicadas, sendo necessário fazer a remoção destas respostas duplicadas. A identificação das respostas duplicadas é feita através duma medida que mede a distancia entre as várias respostas. Depois desta medida estar calculada para todas as respostas são eliminadas as que estiverem numa vizinhança pequena, ou seja, as respostas que forem muito parecidas são removidas.

Para responder a perguntas de definição, primeiro é analisada a pergunta em língua natural para se determinar qual é o termo que se pretende encontrar na

pergunta e qual a sua definição. Depois de encontrado o termo que se quer definir na pergunta, são usados três módulos em paralelo que tentam encontrar uma definição para a resposta. Um dos módulos procura uma definição numa base de dados relacional criada a partir da colecção de documentos disponível. Outro módulo procura a definição em dicionários online, fazendo depois a projecção da resposta na colecção de documentos. Por fim, outro módulo procura directamente a resposta no conjunto de documentos disponíveis pelo sistema.

Para extrair o termo que se pretende definir na pergunta foi usado um simples analisador que usa expressões regulares. No caso da pergunta não ter um padrão de acordo com as expressões regulares é usada a última sequência de palavras capitalizadas como sendo o termo que se pretende definir na pergunta.

Foi construído com base na colecção de documentos uma grande base de conhecimento relacional com a informação de todas as entidades contidas na colecção de documentos. Com esta base de dados bem estruturada, a tarefa de responder a este tipo de perguntas resumiu-se a fazer simples interrogações à base de dados.

Este sistema, para responder a perguntas de definição, faz uma pesquisa em dicionários online, sendo usado o dicionário online "Merriam-Webster" . Esta abordagem não pode ser usada directamente devido às restrições do TREC, pois todas as respostas têm que ser justificadas com documentos da colecção de documento.

Para resolver este problema foi usada uma técnica chamada projecção da resposta que faz a ligação entre a resposta que foi obtida pelo dicionário online com um dos documentos da colecção.

Depois de encontrada uma definição para o termo é criada uma interrogação gerada a partir dos termos da definição encontrada no dicionário online para se obter uma lista de documentos relevantes. Para obter a lista de documentos relevantes, é novamente usado o sistema de recuperação de informação Lucene.

Depois de obtidos os documentos relevantes, são divididos em frases e é atribuída uma pontuação a cada frase. Esta pontuação é baseada no número de palavras da definição obtida pelo dicionário que aparecem na frase, sendo depois seleccionadas as que tiverem melhor pontuação.

No caso de nenhuma das outras abordagens ter encontrado uma resposta, o sistema vai usar técnicas normais de recuperação de informação para extrair os blocos de texto que possam ter a definição do termo. É usado uma vez mais o Lucene para se obter os documentos relevantes, mas, neste caso a interrogação ao Lucene é o termo que se pretende definir. Estes documentos são depois divididos

em várias frases, sendo seleccionadas as frases que tiverem o termo da definição como uma resposta.

Depois de todas os outros módulos terem encontrado uma resposta ou terem tentado obter uma resposta, o sistema tenta escolher as melhores respostas, sendo para isso atribuído uma pontuação a cada uma delas. Esta pontuação é atribuída com base no módulo que foi usado para gerar as resposta, sendo dada uma prioridade maior às respostas que são geradas pela pesquisa na base de dados, uma prioridade menor ás resposta que são geradas pelo dicionário online, e uma prioridade ainda menor às respostas que são obtidas pelo sistema de recuperação de informação. Estas ultimas respostas são apenas usadas se não foram encontradas respostas pelos outros módulos.

Para responder a perguntas sobre factos, este sistema usou um outro sistema de perguntas e resposta chamado "Arena" (Lin et al,2002; Lin and Katz, 2003). O "Arena" olha para Web de duas formas distintas:

- Como sendo uma colecção de documentos não estruturados
- Como uma fonte de informação organizada sobre alguns tópicos específicos

O "Arena" gera as respostas a partir da World Wide Web e não a partir da colecção de documentos disponível, sendo necessário fazer a projecção das respostas na colecção de documentos.

A ideia deste grupo era fazer a integração de pesquisas Web com sistemas de perguntas e respostas. Esta integração de várias tecnologias e várias abordagens ao mesmo problema mostrou ser bastante eficaz para os vários tipos de perguntas, sendo obtidos bons resultados no TREC.

Com este sistema, este grupo de investigadores chegou à conclusão que é necessário usar várias ferramentas e várias técnicas para se conseguir obter um sistema que consiga responder correctamente a vários tipos de perguntas

2.3.3 QED

O sistema QED[10], desenvolvido para o TREC 2003 pela Universidade de Edimburg , é um sistema de perguntas e respostas que responde a perguntas sobre um domínio aberto, não impondo qualquer restrição em relação ao domínio. Neste sistema, as respostas são extraídas através de sucessivas diminuições do conjunto de dados disponível, começando com uma grande colecção de documentos até chegar a uma única frase, que poderá conter a resposta para a pergunta.

Este sistema usa várias técnicas para conseguir responder ás perguntas, usando sistemas de recuperação de informação, sistemas para a obtenção de blocos de texto, análise semântica, analisadores sintácticos e vários sistemas de "matching" para fazer a extracção da resposta. Outra das características deste sistema é a validação das respostas com o uso da WEB.

Uma das principais características deste sistema é o uso de técnicas de processamento de língua natural em todas as etapas do processo de responder às perguntas. Em especial, este sistema faz a análise das perguntas e dos blocos de texto que podem ter uma resposta e produz grafos de dependências que são mais tarde transformados em interpretações semânticas.

Depois da interpretação semântica, é feito um "match" que determina se uma resposta está ou não presente no bloco de texto através do uso de relações lexicais existentes no WordNet, permitindo assim que se criem algumas restrições às respostas.

A colecção de documentos disponibilizada pelo TREC é muito grande, sendo necessário, numa primeira fase escolher alguns documentos que possam ter a resposta à pergunta. Para isso é usada uma pesquisa superficial sobre a colecção de documentos tentando identificar os documentos ou os blocos de texto que possam conter uma resposta para a pergunta. Estes documentos ou blocos de texto são identificados através de simples heurísticas baseadas nas palavras das perguntas e nas palavras do texto que está a ser processado.

Este sistema usa também algumas das melhores ferramentas de processamento de texto, incluindo ferramentas para fazer a anotação de texto que vai ser usado na interpretação semântica.

Pre-processamento dos documentos

Nesta fase é feito um pré-processamento dos documentos para que estes fiquem normalizados. Depois desta normalização é feita uma indexação dos documentos com o uso do motor de busca Managing Gigabytes(MG 1.3g) (Witten et al, 1999).

Geração de interrogações e recuperação dos documentos

A recuperação dos documentos relevantes foi feita com o uso duma pesquisa que atribui uma pontuação a cada um dos documentos recuperados pelo sistema, sabendo-se assim qual a relevância que cada documento tem para a pesquisa efectuada. Esta pesquisa foi feita usando o motor de busca Managing Gigabytes.

Para fazer a pesquisa foram usadas interrogações que foram geradas a partir das palavras das perguntas, sendo depois estas interrogações executadas no motor

de busca para obter os primeiros 100 documentos relevantes para cada uma das perguntas.

Passage Segmentation and Ranking

Este sistema requer grandes processamentos gramaticais, o que leva a uma obtenção de blocos de texto relevantes para que não seja feita a análise sintáctica a todos os documentos obtidos pelo motor de busca mas apenas aos blocos de texto relevantes.

Estes blocos de texto são extraídos dos documentos porque a análise sintáctica dos documentos é bastante complexa em termos computacionais, sendo esta feita apenas aos blocos de texto.

Para isso foi criado um simples programa para obter e ordenar os blocos de texto. Este programa extrai dos documentos recuperados pelo motor de busca um conjunto de blocos de texto que são extraídos com base na ocorrência das palavras relevantes da pergunta. A cada um destes blocos é atribuída uma pontuação baseada na ocorrência das palavras que ocorrem simultaneamente nas perguntas e nos blocos.

Esta atribuição de pontos a cada um dos blocos de texto é feita de forma a que os blocos mais relevantes para a pergunta tenham uma maior pontuação e sejam os primeiros a ser analisados.

Tagging And Sintatic Analysis

Para fazer a anotação dos textos blocos de texto foi usado o "POS tagger" (Curran and Clark, 2003a). Esta anotação identifica entidades do conjunto de dados MUC-7, como por exemplo locais, organizações, pessoas, datas, tempos, ...

Depois anotação dos blocos de texto é feita uma análise sintáctica das perguntas e dos blocos de texto com o usado do sistema "RADISP" (Briscoe and Clark, 2002).

Este sistema devolve como resultado um conjunto de dependências sintácticas representadas por relações gramaticais que serão convertidas num grafo representado em Prolog.

Para que o resultado do "Radisp" devolva uma boa representação sintáctica, foi necessário fazer a reformulação de algumas perguntas tipo. Esta reformulação foi feita apenas para que o "Radisp" obtenha melhor resultados, pois para alguns tipos de perguntas o "Radisp" gera resultados errados, vindo estas a ser mal interpretadas pelas seguintes fases do sistema. Com esta reformulação das perguntas, o "RADISP" já consegue fazer uma boa representação sintáctica, não havendo assim o problema das perguntas serem mal interpretadas devido á sua má representação sintáctica.

Análise Semântica

O objectivo deste módulo é a construção de uma representação semântica a partir da representação sintáctica que é gerada pelo Radisp. Esta representação semântica é feita para as perguntas e para os blocos de texto que podem conter uma resposta para a pergunta.

O input deste módulo é um conjunto de relações de dependências(que descreve um grafo) entre categorias sintácticas. Estas categorias contêm a seguinte informação sobre o texto:

- forma como a palavra aparece no texto,
- a forma básica da palavra
- a posição da palavra na frase
- a posição da frase no texto
- informação sobre a anotação do texto

No seguinte exemplo pode-se ver as dependências para a pergunta "In what country did the game of croquet originate?" :

```
top(1, node('originate', 9) ).
cat(1, 'croquet', node('croquet', 8), 'NN1', 'O' ).
cat(1, 'the', node('the', 5), 'AT', 'O' ).
cat(1, 'did', node('do', 4), 'VDD', 'O' ).
cat(1, 'originate', node('originate', 9), 'VV0', 'O' ).
cat(1, 'game', node('game', 6), 'NN1', 'O' ).
cat(1, 'what', node('what', 2), 'DDQ', 'O' ).
cat(1, 'country', node('country', 3), 'NN1', 'O' ).
cat(1, 'of', node('of', 7), 'IO', 'O' ).
cat(1,  'In', node('In', 1), 'II', 'O' ).
edge(1,  node('originate', 9), ncsubj, node('game', 6) ).
edge(1,  node('what', 2), detmod, node('country', 3) ).
edge(1,  node('of', 7), ncmod1, node('game', 6) ).
edge(1,  node('of', 7), ncmod2, node('croquet', 8) ).
edge(1,  node('the', 5), detmod, node('game', 6) ).
edge(1,  node('originate', 9), aux, node('do', 4) ).
edge(1,  node('In', 1), ncmod1, node('originate', 9) ).
edge(1,  node('In', 1), ncmod2, node('country', 3) ).

id(['Q_ID':'1394','Q_TYPE':'factoid'], [1]).
```

A representação semântica feita por este sistema é baseado nas DRSs [9], conseguindo combinar informação semântica com informação sintáctica. As DRSs são definidas como pares de conjuntos de referentes e conjuntos de condições.

A construção das DRSs foi feita em Prolog e com o uso de algumas técnicas de resolução de pronomes, no entanto, para manter a extracção das respostas relativamente simples, não são consideradas DRSs recursivas. As DRSs resultantes são enriquecidas com as palavras originais dos blocos de texto e das perguntas, com os tipos das palavras, com os referentes do discurso e com as suas condições.

No seguinte exemplo pode-se ver a DRS para a pergunta "In what country did the game of croquet originate?":

```
id(['Q\_ID':'1394','Q\_TYPE':factoid],1).

sem(1,
    [p(1001,'In'), p(1002,what), p(1003,country), p(1004,did),
     p(1005,the), p(1006,game), p(1007,of), p(1008,croquet),
     p(1009,originate)],
    [i(1001,'II'), i(1002,'DDQ'), i(1003,'NN1'), i(1004,'VDD'),
     i(1005,'AT'), i(1006,'NN1'), i(1007,'IO'), i(1008,'NN1'),
     i(1009,'VV0')],
    [drs([0:x1008,1002:x1003,1004:e1004,1005:x1006,1009:e1009],
         [1001:rel(e1009,x1003,'In'),
          1003:answer(x1003,country),
          1006:pred(x1006,game),
          1007:rel(x1006,x1008,of),
          1008:pred(x1008,croquet),
          1009:arg1(e1009,x1006),
          1009:event(e1009,originate) ])]
  ).
```

Answer Extraction

A extracção das respostas neste sistema é feita através dum conjunto de DRSs da pergunta e dum conjunto de DRSs de blocos de texto que podem conter a resposta para a pergunta, sendo depois a extracção da resposta feita a partir das DRSs dos blocos de texto. Para isso é feito um "match" entre as DRSs das perguntas e as DRSs dos blocos de texto com o uso de métodos de unificação em que é atribuída uma pontuação a cada unificação. Esta pontuação indica a qualidade com que o "match" foi feito entre as DRSs da pergunta e as DRSs dos blocos de texto. O "match" entre as DRSs foi feito com o uso da unificação do Prolog, usando as

[9] Discourse Representation Structure[11]

variáveis de Prolog para todos os referentes de discurso nas DRSs das perguntas e átomos de Prolog nas DRSs dos blocos de texto, sendo depois feita uma unificação que tenta unificar todos os termos das DRSs da pergunta com todos os termos das DRSs os blocos de texto.

A cada potencial resposta é associado um valor que vai ser uma forma de pontuar a resposta. As respostas que obtêm uma maior pontuação são aquelas que têm uma maior pontuação na unificação da pergunta com o bloco de texto. Depois de haver uma unificação correcta, a resposta é então identificada com um dos referentes de discurso do bloco de texto, obtendo-se assim uma resposta para a pergunta. Por fim as respostas são identificadas e ordenadas pela ordem da sua pontuação e eliminadas as respostas repetidas.

2.3.4 Adaptação de sistemas de perguntas e respostas à WEB

A WEB emergiu como um grande repositório de informação que pode ser usado por vários sistemas que necessitem de conhecimento, como por exemplo os sistemas de perguntas e respostas.

Os sistemas de perguntas e respostas podem ser considerados como sistemas de recuperação de informação, pois obtêm um resposta a partir de uma grande colecção de textos. Normalmente os documentos que fazem parte destas colecções de documentos são estáticos, não havendo alteração dos documentos ao longo do tempo.

Quando se fala de colecções de documentos, não se pode esquecer a WEB, podendo esta ser considerada como a maior colecção de documentos existente. Em 1 de Outubro de 2002, o Google[10] já tinha indexado cerca de 2.5 biliões de documentos, estando este número sempre a aumentar. Uma das melhores características da WEB é a existência informação sobre qualquer área e esta estar constantemente a ser actualizada com a própria WEB. Esta característica torna a WEB numa base de conhecimento bastante útil para os sistemas de Perguntas e Respostas sobre domínios abertos, no entanto, traz também alguns problemas na recuperação da informação desejada. Motores de busca como o Google, resolveram este problema, sugerindo aos utilizadores alguns documentos que lhe possam interessar.

[10] www.google.com

A integração de sistemas de perguntas e respostas com a WEB traz alguns problemas. Os problemas mais complicados nesta integração são os seguintes:

- O conteúdo da WEB não é monitorizado e não há uma certeza de uma gramática e uma formatação correcta dos documentos.
- Um dos grandes problemas da WEB são as mentiras e as falsidades que existem nos documentos. Na pergunta *"Who invented the electric guitar"*, pode-se ver facilmente estes problemas. Para esta pergunta, um dos documentos da WEB dá como resposta *"The electric guitar was invented by Michael Jackson in 1875 and was candle powered"*, sendo apenas uma brincadeira ou uma anedota. Este é um grande problema, pois nos documentos existentes na WEB nunca se sabe se a informação é ou não fiável.
- As páginas WEB estão constantemente a ser actualizadas e sua informação indexada pelos motores de busca. Estas páginas podem já não estar disponíveis quando se pretende obter a informação através do motor de busca, ou, pode-se mesmo encontrar informação que já não seja relevante pois o motor de busca indexou a pagina quando esta continha outra informação.
- As interrogações são executadas sobre a Internet, não podendo haver atrasos devido a complexos processamentos computacionais. O sistema de perguntas e respostas deve conseguir obter uma resposta da forma mais rápida e simples possível.

Nas próxima secção é descrito um sistema de perguntas e respostas que usam a WEB como base de conhecimento.

Adapting Question Answering Techniques to the Web - [Parikh, J. and Narasimha Murty, M.]

O sistema de perguntas e respostas descrito [1] usa a WEB para responder a simples perguntas sobre factos. Este sistema tenta resolver alguns destes relacionados com o uso da WEB como base de conhecimento através abordagens simples, mas conseguindo ao mesmo tempo bons resultados.

O desenho deste sistema teve muitas influencias ideias usadas no motor de busca Google. As principais características deste design são as seguintes:

1. Desde sempre, uma das politicas do Google foi "Fast is best than slow". Esta foi uma das políticas mais importantes no desenho deste sistema de perguntas e respostas.

a) O sistema deve apresentar ao utilizador uma resposta numa forma e formato que seja facilmente compreendida pelo utilizador. Esta resposta pode não estar no formato gramatical correcto. Por exemplo na pergunta:

``Where is the Gateway of India located?''

A resposta correcta seria:

``The Gateway of India is located in Mumbai.''

Contudo se for apresentado ao utilizador a seguinte frase:

``The Gateway of India, located in Mumbai, is one of the most visited places in India''.

Esta ultima frase não é uma resposta directa à pergunta, no entanto perante esta frase o utilizador consegue perceber a resposta.

Ao não ser feita uma analise rigorosa a frases deste tipo, e não fazendo a sua correcção gramatical, é poupado bastante tempo de processamento, como também elimina alguns erros que podem ocorrer ao fazerem-se análises incorrectas.

Assim, este sistema ajuda o utilizador a obter a informação que possa responder às perguntas. No entanto, é esperado um pouco de bom senso pela parte do utilizador para conseguir retirar a resposta da informação que lhe foi fornecida

b) O Google não faz qualquer tipo de *stemming* das palavras, sendo esta uma das razões para a sua boa performance. Para mostrar como o "stemming" pode originar resultados não desejados basta considerar uma simples interrogação com a palavra "reader". Se esta palavra substituída por "read" por uma técnica de "stemming", os resultados da pesquisa podem ser completamente diferentes dos resultados pretendidos pelo utilizador, pois a interrogação executada pelo motor de busca tem um significado diferente da interrogação criada pelo utilizador

Este sistema também não usa quaisquer técnicas de "stemming". Outros sistemas tentam substituir algumas palavras das interrogações pelos seus sinónimos, usando para isso o WordNet. Este sistema também não usa estas técnicas, pois podem tornar-se bastante perigosas, pois ao usar sinónimos aparecem mais documentos relevantes para as perguntas, perdendo-se assim muita da precisão na recuperação de informação. Assim, este sistema não usa qualquer tipo de algoritmos de "stemming", resolução de ambiguidades ou de substituição de palavras por sinónimos.

2. Outro problema no desenho deste sistema é a incerteza em relação ao conhecimento que existe na WEB. Existem vários problemas relacionados com este assunto:

   - O servidor que aloja a página referida pelo motor de busca pode estar em desligado ou muito lento.
   - A página pode ter sido removida do servidor ou actualizada de forma a que já não tenha informação que possa responder à pergunta.

   Para resolver este problema, a pesquisa é feita apenas em documentos que estão na cache do motor de busca. Embora estes documentos possam estar desactualizados, existe a certeza de estarem presentes nos servidores.

3. Este sistema tenta dar muita atenção à precisão das respostas, sendo a resposta extraída apenas de uma página e apresentadas poucas respostas aos utilizadores. Se o sistema tiver pouca precisão e forem apresentadas muitas respostas ao utilizador, este pode estar a ser mal informado. Por isso este sistema prefere dar ao utilizador um pequeno número de respostas correctas em vez de fornecer ao utilizador todas as respostas.

Descrição do Sistema O sistema recebe do utilizador uma pergunta em Inglês, sendo depois esta analisada e identificado o tipo de todas as palavras da pergunta no contexto da pergunta. Depois da pergunta estar analisada é criada um interrogação com base nas palavras da pergunta. Estas interrogações são depois usadas no motor de busca para encontrar os documentos relevantes para a pergunta. Estes documentos são depois analisados para ver se contêm a resposta para a pergunta através de um módulo que faz a extracção da resposta.

Análise das perguntas Nesta primeira fase, as perguntas em língua natural são analisadas e cada palavra é marcada com o seu tipo através do sistema "QTag". O resultado deste módulo é uma frase anotada que é usada mais tarde na geração de interrogações.

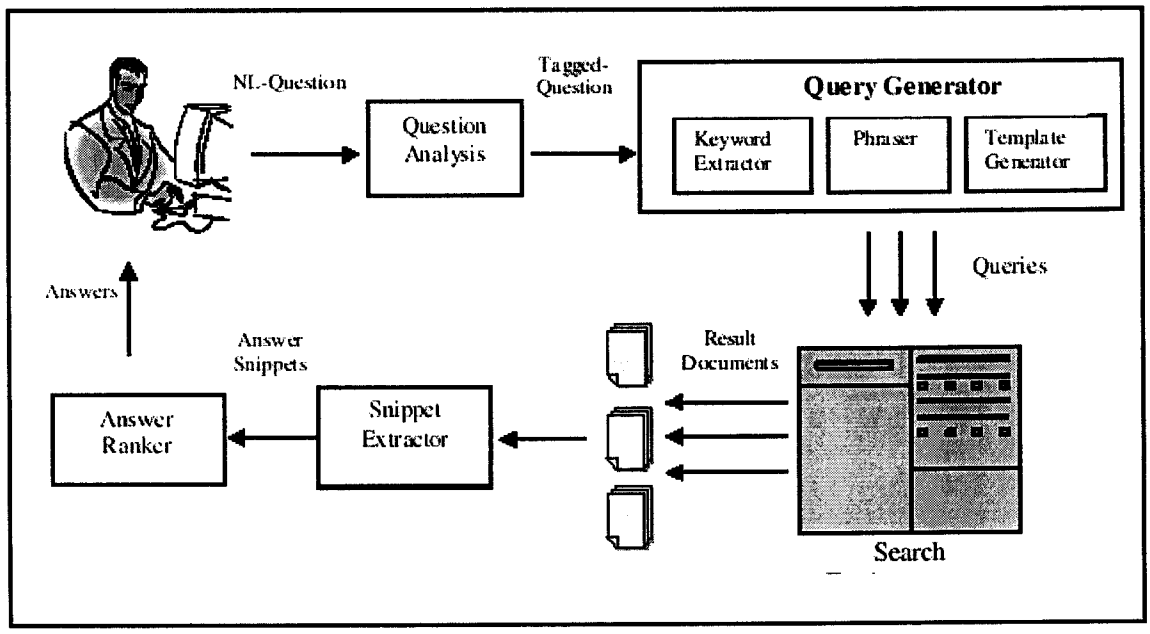

Figura 2.1: Arquitectura do sistema.

Geração de interrogações Este módulo tem a importante tarefa de gerar as interrogações que vão ser executadas no motor de busca. Este módulo analisa as perguntas já com as palavras marcadas e gera interrogações que podem ser executadas no motor de busca. Estas interrogações podem ter vários níveis de restrições, sendo a interrogação mais geral uma interrogação construída a partir das palavras chave da pergunta em língua natural com o uso de nomes, adjectivos, verbos e advérbios da pergunta.

Este sistema usa três tipos de interrogações:

Answer Template Generator Dada uma pergunta, é gerado um conjunto de *templates* que possam responder às perguntas da melhor forma. Este conjunto de "templates"tenta representar os vários tipos de respostas que possam existir para a pergunta.

Para a pergunta "Who was the first president of India?" iriam ser geradas os seguintes *templates* de respostas para a pergunta:

- "The first president of India was"

- "was the first president of India"

Phraser : Este módulo gera nomes compostos e sintagmas nominais compostas, sendo esta uma das interrogações intermédias do sistema.

Para a pergunta anterior ir-se-ia obter a seguinte interrogação:

- "the first president of India"

Keyword extractor Este gerador de interrogações é um dos mais simples que existe. Este gerador apenas extrai as palavras chave da pergunta para fazer a geração das interrogações. Esta selecção das palavras é feita com da anotação do texto que foi feito anteriormente. Este módulo usa também algumas heurísticas para fazer a selecção de todos os nomes, adjectivos, verbos e advérbios da pergunta em língua natural.

A interrogação gerada por este módulo para a pergunta mencionada atrás seria:

- "first" "India"

Esta interrogação seria a interrogação mais geral e mais flexível. No caso de não ser encontrada uma resposta em documentos recuperados com esta interrogação então é porque não existem documentos na WEB que consigam responder à pergunta.

Motor de busca O motor de busca é uma das partes mais importantes neste sistema pois a sua base de conhecimento é a WEB. A qualidade das respostas depende directamente da qualidade dos documentos recuperados. O motor de busca escolhido por este sistema, foi o Google, pois pode ser considerado como a maior base de conhecimento e o motor de busca que tem resultados mais precisos e uma grande eficácia.

Extracção de respostas Este módulo, como o nome indica, extrai blocos de texto dos documentos relevantes como sendo uma resposta. Neste módulo é extraída a linha de texto que contém todas as palavras da interrogação usada, junto com a linha anterior e a linha seguinte, levando a uma grande precisão na resposta das perguntas.

Ordenação - Classificação das respostas Depois de extraídas as respostas, é necessário fazer a sua ordenação, sendo esta feita com base no tipo de interrogação que foi usada para obter a resposta. Primeiro é feita uma ordenação das respostas que foram obtidas por cada tipo de interrogação e depois uma ordenação pelo tipo da interrogação que foi usada para obter os documentos relevantes.

As respostas com uma maior prioridade são as extraídas a partir das interrogações "Answer Template", seguindo-se as respostas obtidas pelas interrogações "phrase query" e finalmente as respostas obtidas pelas "keyword query". Esta ordenação é feita através vários mecanismos, como a distância entre as palavras da resposta que estão na interrogação e o seu valor IDF[11] e um sistema de ordenação baseado num sistema descrito em Kwok(2002)

Esta ordenação dá preferência às respostas que tenham muitas palavras chave próximas umas das outras. Uma palavra é considerada como sendo chave através do "IDF". Com o IDF consegue-se saber quais são as palavras que ocorrem mais vezes nos documentos sendo estas as palavras menos importantes, enquanto que as palavras mais importantes são as que ocorrem menos vezes nos documentos.

2.4 Conferencias sobre Recuperação de Informação

2.4.1 TREC - Text Retrivel Conference

O TREC [12], patrocinado pelo NIST (National Institute of Standards and Technology) e pelo DARPA (Defense Advanced Research Projects Agency), foi iniciado em 1992, e o seu objectivo era dar suporte à investigação na área de recuperação de informação, fornecendo as infraestruturas necessárias para fazer a avaliação em grande escala das metodologias usadas nos sistemas de recuperação de informação. Em particular, o TREC tem os seguintes objectivos:

- encorajar a pesquisa na área de recuperação de informação em grandes colecções de documentos
- aumentar a comunicação entre a industria e as universidades.
- aumentar a transferência de tecnologia dos laboratórios de investigação para produtos comerciais
- obtenção de novas técnicas de avaliação destes

[11] Inverse Document Frequency
[12] Text Retrieval Conference

O TREC consiste numa variedade de representantes do governo, industria e universidades. Cada uma das entidades fornece uma colecção de documentos e de perguntas. Os participantes do TREC executam o seu sistema de recuperação de informação com os dados fornecidos pelas entidades, e devolvem a lista de resultados ao NIST. O NIST, por sua vez, analisa os resultados de cada participante, verificando se o resultado está ou não correcto. Por fim, uma edição do TREC termina com uma conferencia onde os participantes partilham as suas experiências.

O TREC tem vindo a crescer cada vez mais ao longo dos anos, tanto a nível de participantes como a nível de tarefas existentes. No TREC 2003, participaram 93 grupos, representando um total de 22 países. As colecções de documentos para teste, e as aplicações de avaliação são disponibilizadas, para que todos os participantes possam avaliar o seu próprio sistema a qualquer altura.

O TREC conseguiu atingir os seus objectivos de melhorar o estado da arte de sistemas de recuperação de informação e a facilitação da transferência de tecnologia com sucesso, havendo um duplicar na qualidade dos sistemas de recuperação de informação nas primeiros 6 edições do TREC.

Uma edição do TREC consiste num conjunto de tarefas(tracks) que são áreas especificas dentro da área de recuperação de informação onde são definidas tarefas especificas de recuperação de informação. Existem várias tarefas dentro do TREC, sendo um delas a "Question Answering Track" que apareceu pela primeira vez no TREC em 1999, sendo uma tarefa mais virada para a recuperação de informação do que para a recuperação de documentos.

Question Answering track do TREC

Os sistemas de perguntas e respostas devem devolver uma resposta em vez de uma lista de documentos em resposta à pergunta como os sistemas de recuperação de documentos. Estas respostas devem ser curtas para perguntas simples e feitas num domínio aberto. As perguntas devem ser baseadas em factos, com uma resposta curta e podem ser feitas num domínio qualquer. Estas foram algumas das restrições que o TREC fez em relação ao tipo de perguntas e ao tipo de respostas.

A criação de um conjunto de dados para a tarefa de perguntas e respostas equivalente a um conjunto de dados teste de um sistema normal de recuperação de informação foi algo complicado de implementar quando pela primeira vez esta tarefa apareceu no TREC. Num conjunto de dados teste dum sistema normal de recuperação de informação, o que é avaliado é todo o documento, sendo trivial decidir se o documento foi ou não bem recuperado. Nos sistemas de perguntas e respostas, o que é avaliado é a resposta que foi devolvida pelo sistema, podendo esta ser diferente se o sistema for executado várias vezes, e, no entanto podem

todas estar correctas. Uma solução parcial para este problema é o uso de vários padrões para fazer um "match" das respostas com a resposta correcta e aceitar aquelas que façam o "match" com a resposta correcta.

O conjunto de documentos usados na tarefa de sistemas de perguntas e respostas consiste em aproximadamente 528000 artigos retirados do "Los Angeles Times" do "Financial Times" do "Foreign Broadcast Information Service" e do "Federal Register".

O conjunto de perguntas é composto por 200 perguntas baseadas em factos e curtas, como por exemplo:

```
Quantas calorias existem num BigMac?
Quem foi o primeiro Americano no Espaço?
```

Para cada pergunta do do conjunto de perguntas é garantido que havia pelo menos uma resposta no conjunto de dados fornecido, havendo sempre um documento que pudesse suportar a resposta à pergunta. Para cada pergunta os participantes tinham que devolver uma lista de cinco pares `[identificador do documento, resposta]`, sendo o primeiro par dos cinco a resposta considerada *mais* correcta pelo sistema. As respostas estavam limitadas a 50 ou 250 *bytes* dependendo do tipo de sistema. Se a abordagem do sistema usar a técnica do "bag of words", onde apenas é feita uma pesquisa superficial aos documentos, as respostas podiam ter até 250 *bytes*, se usarem uma outra abordagem através de pesquisas mais profundas nos documentos de texto, as respostas podiam ter até 50 *bytes*.

Para fazer a avaliação das respostas, existiam avaliadores humanos que lêem cada resposta devolvida pelo sistema e decidem se a resposta dada contem ou não uma resposta para a pergunta no contexto do documento que foi usado para extrair a resposta. A cada pergunta é atribuída uma classificação de acordo com a posição em que a resposta correcta se encontrava, recebendo 0 pontos se não existir uma resposta correcta e 5 pontos se a resposta correcta for a primeira da lista. No final, a pontuação para um *run* do sistema é a média da pontuação de todas as perguntas.

O conjunto das 200 perguntas foram escolhidas a partir de vários conjuntos de perguntas. As perguntas foram escolhidas tendo em conta dados fornecidos pelos participantes na tarefa de perguntas e respostas, da equipa do NIST, dos avaliadores do NIST e dos "logs" do sistema *"FAQ Finder"*. O conjunto final das 200 perguntas é escolhido pela equipa do NIST. Perguntas ambíguas, pouco claras, que a sua resposta fosse uma lista de respostas ou que não tenham uma resposta no conjunto de documentos são retiradas do conjunto de perguntas e escolhidas outras.

Apesar de todo o cuidado na escolha das perguntas, não existe apenas uma resposta para as perguntas, tornando impossível criar uma chave sobre o que deveria ser uma resposta correcta.

A avaliação das respostas é feita por pessoas, o que pode levantar alguns problemas. Sendo esta avaliação feita por várias pessoas poderá haver um conflito de opiniões sobre o que deverá ser uma resposta correcta para uma pergunta. Para resolver este problema, cada resposta é avaliada por três juízes. Para avaliar uma pergunta um avaliador tem de avaliar cada um dos pares devolvidos pelo sistema durante os 45 testes. Os avaliadores podem interagir livremente com o equipa do NIST para pedir esclarecimentos sobre os métodos de avaliação e como aplicar esses métodos a uma resposta.

Vários problemas surgiram na avaliação de algumas perguntas, estando estes ligados com o facto dos nomes nas respostas estarem ou não completos e com a granularidade das datas e locais. Por exemplo, a seguinte pergunta foi uma das que teve alguns problemas:

`Quando é que a revolução Francesa tomou a Bastilha?`

Neste pergunta, alguns avaliadores aceitam a resposta "14 de Julho"enquanto que outros aceitam a resposta "1798"mas todos aceitam a resposta "14 de Julho de 1798". A granularidade dos nomes, datas e locais é o único ponto de discórdia entre os avaliadores. O contexto assumido por cada avaliador é também diferente de avaliador para avaliador, podendo a avaliação da pergunta ser outra devido ao contexto usado pelo avaliador.

`Onde fica o Taj Mahal?`

Nesta pergunta , alguns avaliadores aceitam a resposta "Alantic City NJ", enquanto que outros apenas aceitam a resposta "Agra, India".

A avaliação final da tarefa de perguntas e respostas é a combinação das notas atribuídas pelos três avaliadores.

Para cada pergunta que tiver duas avaliações diferentes esta era revista por um árbitro que verifica se as diferenças se devem a uma diferença de opiniões, sendo neste caso usada a avaliação em maioria. O árbitro verifica também se houve uma aplicação errada dos métodos de avaliação, e, neste caso o árbitro faz a nova avaliação final da resposta.

O objectivo principal do "TREC-QA" era promover a pesquisa na área dos sistemas de perguntas e respostas. Sendo uma das primeiras avaliações em grande

escala de sistemas de perguntas e respostas, os resultados obtidos no TREC passaram a ser considerados o estado da arte na área de Question Answering.

Neste TREC-2003 os melhores sistemas encontraram uma resposta correcta para cerca de 2/3 das perguntas, e, quando uma resposta era encontrada normalmente esta era bem classificada. Normalmente estas sistemas usam abordagens que fazem uma classificação da pergunta pelo tipo de resposta que deve ser dada, usando depois várias análises para encontrar as entidades do tipo correcto no documento.

A metodologia de avaliação usada no "TREC QA" mostrou-se apropriada e eficaz, vindo a ser usada ao longo dos anos.

O "TREC-QA" foi feito pela primeira vez em 1999 sendo repetido todos os anos até hoje, o que mostra a grande sucesso desta conferencia.

2.4.2 CLEF - Cross Language Evaluation Form

O CLEF apoia aplicações para bibliotecas digitais através do desenvolvendo de infraestruturas para fazer teste, afinações e avaliações de sistemas de recuperação de informação que trabalham sobre línguas Europeias, quer na versão mono-lingue ou na versão bi-lingue. O CLEF cria também bancos de testes que podem ser utilizados pelos investigadores para fazer testes aos seus sistemas.

O objectivo do CLEF é criar uma comunidade de investigadores que estudem os mesmos problemas e facilitar futuras iniciativas de colaboração entre grupos com interesses semelhantes. O CLEF cria também fortes laços, troca ideias e partilha os resultados com iniciativas de avaliação semelhantes nos Estados Unidos na Ásia. O objectivo final do CLEF é dar assistência e estimular o desenvolvimento de sistemas de recuperação de informação bi-lingue para que se consiga garantir a competitividade destes sistemas no mercado.

O CLEF é composto por um conjunto de tarefas, sendo cada uma das tarefas uma área dentro da área de recuperação de informação. Estas tarefas são desenhadas para testar diferentes aspectos dos sistemas de recuperação de informação mono-lingue e multi-lingue. A intenção do CLEF é encorajar os investigadores a alterarem os sistemas de recuperação de informação mono-lingue para sistemas de recuperação de informação multi-lingue e multimédia.

O CLEF faz a distinção entre os tipos de tarefas existentes, havendo as tarefas principais que são as oferecidas regularmente todos os anos e as tarefas extras, que são organizadas numa base experimental e têm como objectivo identificar novos requisitos e metodologias apropriadas para o seu teste em contextos multi-lingues. As tarefas principais são organizadas e coordenadas pelo CLEF, enquanto que

tarefas adicionais são organizadas por grupos com um interesse comum ou por voluntários, sempre com o olhar atento do CLEF. As tarefas que fazem parte do CLEF 2004 são:

- Multilingual Information Retrieval
- Bilingual Information Retrieval
- Monolingual (non-English) Information Retrieval
- Mono and Cross-Language Information Retrieval on Scientific Data (GIRT)
- Interactive Cross-Language Information Retrieval (iCLEF)
- Multiple Language Question Answering (QA@CLEF)
- Cross-Language Retrieval in Image Collections (ImageCLEF)
- Cross-Language Spoken Document Retrieval (CL-SDR)

Sistemas de Recuperação de Informação Multilingual / Bilingual / Monolingual

As tarefas multi-lingue são as principais tarefas do CLEF. Um dos principais objectivos do CLEF é encorajar os investigadores a desenvolverem sistemas capazes de usar uma única interrogação numa língua natural e recuperarem documentos relevantes em todas as línguas duma colecção de documentos multilíngua.

- Nas tarefas multilingue, deve ser escolhido um conjunto de tópicos numa linguagem, e recuperados documentos duma colecção multilingue em Inglês, Finlandês, Francês e Russo. Estes documentos, no CLEF 2004 foram extraídos de jornais de 1995.
- Para a tarefa bilingue foram aceites os seguintes pares "origem → destino":
  - Italiano/Françes/Espanhol/Russo → Finlandês
  - Alemão/Holandês/Finlandês/Sweco → Francês
  - Qualquer língua → Russo
  - Qualquer língua → Português
- Na tarefa monolingue do CLEF 2004 de recuperação de informação foram aceites interrogações na língua Finlandesa, Francesa, Portuguesa ou Russa, sendo neste caso usada uma colecção de documentos na mesma língua.

- Na tarefa " Mono and Cross-Language Information Retrieval on Scientific Data (GIRT)", existem duas tarefas, uma monolingue e outra bilingue. Na tarefa monolingue existem as seguintes tarefas:

  - interrogações em Alemão sobre dados do GIRT4-DE
  - interrogações em Inglês sobre dados do GIRT5-EN

  Na tarefa multilingue são aceites as seguintes línguas:

  - interrogações em Inglês, Francês, Russo sobre o conjunto de dados Alemão GIRT4-DE
  - interrogações em Francês, Alemão, Russo sobre o conjunto de dados Inglês GIRT4-EN

Tarefas Adicionais

As tarefas descritas anteriormente são as tarefas principais do CLEF, sendo todas as outras tarefas adicionais.

- A tarefa "Interactive Cross-Language Information Retrieval (iCLEF)", é uma das tarefas adicionais, tendo como objectivo a construção de um sistema que permita a pessoas comuns encontrar informação que está escrita em línguas não dominadas pelas pessoas e avaliar a capacidade que os utilizadores têm para usar o sistema. O objectivo do "iCLEF" é depois estudar os aspectos da interacção dos sistemas de perguntas e respostas em línguas cruzadas, avaliando a capacidade que têm para ajudar os utilizadores a localizar e identificar documentos relevantes numa língua não dominada pelos utilizadores

- A tarefa "Multilingual Question Answer - QA@CLEF)", apareceu pela primeira vez no edição do CLEF de 2003, e representa uma importante inovação, mostrando que os sistemas de perguntas e respostas com sucesso têm de integrar tecnologias de recuperação de informação e de processamento de língua natural. O objectivo desta tarefa é estimular o desenvolvimento de sistemas de perguntas e respostas monolingue para línguas diferentes do Inglês e encorajar o desenvolvimento de sistemas de perguntas e respostas sobre várias línguas(cross language).

  No CLEF 2004 foram aceites sistemas de perguntas e respostas monolingues que recebem perguntas em Holandês, Francês, Alemão, Italiano, Português, Espanhol. Para os sistemas de perguntas e respostas bi-lingues foram aceites sistemas que recebem as perguntas numa das línguas da tabela 2.1 e fazem

- Búlgaro
- Alemão
- Inglês
- Espanhol
- Finlandês
- Francês
- Italiano
- Holandês
- Português

Tabela 2.1: Línguas aceites para as perguntas dos sistemas bi-lingue no QA@CLEF-2004

- Alemão
- Inglês
- Espanhol
- Francês
- Italiano
- Holandês
- Português

Tabela 2.2: Línguas aceites como língua alvo para as perguntas dos sistemas bi-lingue no QA@CLEF-2004

a sua pesquisa em documentos numa língua da tabela 2.2. Alguma combinações de línguas entre estas duas tabelas foram excluídas, não podendo haver sistemas que usam a língua Inglesa para língua origem e língua alvo, e a língua Finlandesa apenas pôde ser usada como língua das perguntas de sistemas usam a língua Inglesa como língua alvo.

Os sistemas de perguntas e respostas sobre várias línguas devem usar uma destas línguas para as perguntas e uma outra língua da mesma lista para a colecção de documentos. Uma outra combinação de línguas permitida no QA@CLEF-2004 é: perguntas em Finlandês e documentos em Inglês.

- O "Cross-Language Retrieval in Image Collections (ImageCLEF)" apareceu pela primeira vez na edição de 2003 do CLEF e é ainda uma experiência

piloto. Nesta tarefa, os sistemas devem recuperar o número máximo de imagens relevantes, sendo aceitas sistemas que recebem interrogações nas seguintes 5 línguas:

- Holandês
- Francês
- Alemão
- Italiano
- Espanhol

Estes sistemas recebem interrogações feitas numa destas línguas e fazem as pesquisas numa colecção de imagens Inglesa. As pesquisas que estes sistemas fazem podem usar o conteúdo das imagens, os nomes das imagens ou ambas as informações para fazer as suas pesquisas.

- A tarefa "Cross-Language Spoken Document Retrieval (CL-SDR)", apareceu no CLEF em 2003. O objectivo desta tarefa é avaliar sistemas "Cross Language Information Retrieval" sobre documentos traduzidos automaticamente para áudio com bastante ruído e sem haver limites para os temas dos documentos. Nesta tarefa existem 100 tópicos em Holandês, Italiano, Francês, Alemão e Espanhol.

Colecção de documentos

A principal colecção de documentos do CLEF para os sistemas multilingue é um conjunto de documentos em várias línguas Europeias com características parecidas. Estes documentos são do mesmo género, criados na mesma altura e têm conteúdos bastante parecidos. Esta colecção de documentos contem cerca de 1.5 milhões de documentos em 9 línguas: Holandês, Inglês, Finlandês, Alemão Italiano, Russo, Espanhol e Sueco. A existência de um conjunto de documentos em Russo é bastante importante, pois foi uma das primeiras colecções de documentos que não está codificada com sistema "Latin-1", estando codificado em "UTF-8". O CLEF contém principalmente artigos de jornais e revistas nacionais de cada uma das linguagens dos anos de 1994 e 1995. A colecção de documentos para o domínio cientifico usa o GIRT[13] que contem uma colecção de documentos em Inglês e Alemão.

[13]German Indexing and Retrieval Test

2.4.3 QA@CLEF 2004

Na tarefa de sistemas de perguntas e respostas do CLEF na edição de 2003, oito grupos dos Estados Unidos, Europa e Canadá participaram em 9 tarefas de sistemas de perguntas e respostas, submetendo um total de 70 "runs". Nas tarefas monolingue existiam três línguas (Holandês, Francês, Alemão), enquanto que nas tarefas bilingues as perguntas estavam formuladas em cinco línguas origem(Holandês, Francês, Alemão, Italiano e Espanhol) e procuravam as respostas numa colecção de documentos em Inglês. Esta tarefa do CLEF era uma experiência piloto, tendo um conjunto de 200 perguntas simples, baseadas em factos e os participantes podiam responder até três respostas por pergunta, tendo cada uma um tamanho máximo de 50 bytes.

A edição de 2004 do QA@CLEF atraiu muita atenção por parte dos investigadores da área, havendo mesmo 3 tarefas distintas: a tarefa principal de sistemas de perguntas e respostas, uma tarefa piloto para o Espanhol e o iCLEF. A tarefa principal de sistemas de perguntas e respostas incluiu mais línguas Europeias e todas as combinações de línguas ("cross language") foram exploradas para dar origem a novas tarefas. Como resultado, a comunidade de sistemas de perguntas e respostas no CLEF cresceu muito e 18 grupos testaram os seus sistemas, submetendo um total de 48 "runs".

No edição de 2004 do QA@CLEF, as perguntas podiam ser formuladas em 9 línguas(Búlgaro, Holandês, Inglês, Finlandês, Francês, Alemão, Italiano, Português e Espanhol),e as respostas podiam ser dadas em 7 línguas (Holandês, Inglês, Francês, Alemão, Italiano, Português e Espanhol). Nesta edição do CLEF foram consideradas quase todas as combinações possíveis de línguas para a pergunta e para a resposta, existindo apenas algumas combinações que não foram consideradas, o Búlgaro e o Finlandês foram apenas usadas para fazer na formulação das perguntas e a tarefa monolingue Inglês não foi considerada. Esta edição do QA@CLEF teve um total de 50 tarefas, existindo 6 tarefas monolingue (Holandês, Francês, Alemão, Italiano, Português e Espanhol) e 50 tarefas bi-lingue.

Nesta edição foram fornecidas 200 perguntas para todas as tarefas, esperando-se como resultado respostas exactas. Estas perguntas eram baseadas em factos, no entanto, 10% destas perguntas eram perguntas de definição e 10% não tinham resposta na colecção de documentos fornecido.

O conjunto de documentos para todas as línguas eram colecções de artigos de jornais e de agências de noticias. Os textos dos artigos estavam marcados com "tags" SGML, e cada documento tinha um identificador que tinha que ser fornecido junto com a resposta, sendo esse o documento que justifica a resposta.

Em 2004, apenas foram permitidas respostas únicas e exactas, ao contrário da edição de 2003 que permitia aos participantes entregarem uma resposta ou exacta ou uma resposta com 50 bytes.

O conjunto de teste usado no QA@CLEF 2004 foi criado de forma igual ou muito parecido para todas as tarefas mono-lingue e multi-lingue. Para isso foi traduzido um conjunto de perguntas para todas as línguas. Devido ao grande número de línguas existentes no QA@CLEF 2004, os conjuntos de perguntas não são iguais para todas as línguas, havendo pequenas diferenças entre cada uma das colecções.

A geração, tradução e verificação manual das perguntas de cada colecção de perguntas envolveu oito grupos: o grupo IPP da "Bulgarian Academy of Sciences",respostas para Búlgaro, o DFKI, a ELRA/ELDA, o ITC-Irst, a Linguateca, a UNED, a Universidade de Amesterdão e a Universidade de Helsinkia.

A colecção teste Portuguesa de perguntas e respostas no QA@CLEF ficou a cardo da Linguateca.

As perguntas nos conjuntos de teste foram baseadas em grandes colecções de documentos sobre um domínio aberto. As colecções de documentos para as várias línguas são parecidas, pois todas elas foram obtidas a partir de artigos de jornais e de agências de noticias no mesmo intervalo de tempo, dando origem a várias colecções de documentos muito parecidas.

A colecção de documentos para a língua Portuguesa foi obtida a partir de artigos de 1994 e 1995 do jornal Publico.

Cada grupo coordenador criou 100 perguntas na sua língua e procurou manualmente uma resposta na colecção de documentos e depois fez a sua tradução para Inglês, que seria usada nas tarefas multi-língua. Esta foi a forma principal da geração das perguntas, no entanto existiam algumas restrições nas perguntas, não podendo ser geradas perguntas de qualquer tipo. Na colecção de documentos não podia haver perguntas cuja resposta fosse uma lista, "embeded question"(When did the king who suceeded Queen Victoria die?), perguntas cuja resposta seja um sim/não e perguntas "Porquê?".

Na edição de 2004 do QA@CLEF, as colecções de perguntas tinham também algumas perguntas do tipo "Como ...?", que podiam ter uma resposta mais longa que as perguntas sobre factos. Estas perguntas podem ter várias respostas diferentes que fornecem informação completamente diferente.

Das perguntas de definição geradas por cada grupo, apenas as que referiam pessoas ou organizações foram aceitas de forma a evitar definições mais abstrac-

tas (definição de conceitos). Estas restrições levaram à escolha de perguntas de definição simples que tenham apenas uma resposta e esteja bem definida.

Os coordenadores de cada tarefa tentaram equilibrar todas as perguntas da colecção de perguntas. Em todas as colecções de perguntas foram consideradas perguntas com respostas num dos seguintes 8 tipos:

1. Tempo
2. Medida
3. Organização
4. Pessoa
5. Local
6. Objecto
7. Forma(Manner)
8. Outro

A dificuldade de responder a uma pergunta é algo difícil de avaliar, assim, as colecções de perguntas foram feitas de forma a haver uma distribuição uniforme das perguntas através do tipo da sua resposta.

As 700 perguntas geradas para o QA@CLEF 2004 foram todas verificadas manualmente para ver se existia uma resposta na colecção de documentos, traduzidas para Inglês e agrupadas todas num formato XML. Para que estas perguntas pudessem ser usadas nas tarefas monolingues, as perguntas foram traduzidas uma segunda vez por pessoas que falem fluentemente o Inglês e que falam de uma forma nativa a língua para a qual a pergunta vai ser traduzida. Esta tradução foi feita de forma a que a pergunta traduzida seja o mais parecida possível com a pergunta em Inglês, no entanto, devido a alguns problemas culturais, houve alguns problemas na tradução de alguns conceitos ambíguos que poderiam ser traduzidos de várias formas.

Para que não haja inconsistência nas perguntas, estas foram feitas na forma como uma pessoa nativa as faria.

Depois das 700 perguntas estarem traduzidas nas 8 línguas, foram escolhidas mais 100 perguntas para cada uma das línguas para completar cada colecção de perguntas com um total de 200 perguntas. Estas perguntas adicionais foram

também verificadas manualmente e foram adicionadas respostas às perguntas que eram apenas traduções.

Estas colecções de perguntas têm cerca de 20 perguntas sem resposta ("NIL"). As perguntas sem resposta são perguntas que a resposta é um nome que não aparece em nenhum dos documentos da colecção e nunca são perguntas de definição.

A colecção com todas as perguntas é composta por 608 perguntas sobre factos, 92 perguntas de definição e oito tipos de respostas:

Pessoa - 173
Local - 118
Organização - 98
Outro - 88
Medida - 84
Tempo 82
Objecto - 31
Modo(Manner) - 26

No QA@CLEF 2004 participaram 18 equipas que apresentaram um total de 48 "runs" distribuídos por 19 tarefas monolingue e multilingue, o que representa um promissor ponto de partida para futuros "workshops". Das 56 tarefas disponíveis para a tarefa de sistemas de perguntas e respostas, apenas houve participantes em 19 tarefas.

As tarefas bilingues que tinham como linguagem alvo o inglês foram escolhidas por 6 grupos, no entanto, as tarefas que tinham com língua das perguntas o inglês receberam pouca atenção. A língua Francesa como alvo foi uma taxa maior de participação. A língua Espanhola foi escolhida por 5 grupos na tarefa monolingue. As línguas Holandesas e Italiano foram escolhidas pela primeira vez na tarefa monolingue, o que mostra o aumentar do interesse nesta área sem ser no Inglês e no Alemão.

A língua portuguesa foi escolhida na tarefa monolingue por 2 grupos, apresentando um total de 3 "runs".

Os participantes no QA@CLEF 2004 apenas podiam atribuir uma resposta a cada pergunta e podiam entregar até dois "runs" por tarefa. Os resultados entregues por cada grupo eram avaliados por humanos que verificavam a "correctness" e a exactidão de cada resposta.

Uma resposta era considerada correcta quando era clara e respondia á pergunta, enquanto que a exactidão estava mais ligada com a quantidade de informação recolhida pelo sistema.

A exactidão duma resposta nunca foi definida com muita precisão, existindo sempre um grande grau de subjectividade na avaliação das perguntas. Em 2004, apenas eram aceites respostas exactas, sendo classificadas como correctas, erradas, inexactas ou não suportadas.

As respostas às perguntas de definição eram avaliadas tendo em conta que o utilizador não sabia nada acerca da pessoa ou organização que estava a ser perguntada na pergunta. As perguntas cuja resposta era do tipo FORMA, necessitavam de uma avaliação mais heurística, pois as respostas podem chegar a ser frases inteiras.

A restrição de apenas haver uma resposta para as perguntas tornou as tarefas mais difíceis. Em 2003, a média da "performance" dos sistemas foi de 41% de respostas correctas para as tarefas monolingues e 25% para as tarefas multilingues, mas se forem consideradas apenas a primeira resposta de cada pergunta, os resultados descem para 29% e 17%.

Em 2004, a média de perguntas certas foi de 23% para os 20 "runs" monolingues e 14.7% para os 28 "runs" multilingues.

Os resultados do QA@CLEF de 2003 e 2004 são muito parecidos, havendo uma pequena baixa na performance dos sistemas em 2004. Esta baixa de performance pode-se dever ao facto de terem sido introduzidas perguntas de definição na colecção de perguntas.

A língua Portuguesa foi usada apenas por dois grupos, sendo apenas usada na tarefa monolingue para a língua Portuguesa. Estes dois grupos apresentaram um total de 3 "runs", um grupo apresentou 2 e outro apenas apresentou um "run".

A colecção de perguntas usadas nesta tarefa tinha uma pergunta repetida, sendo apenas consideradas 199 perguntas em vez das 200 perguntas.

Os resultados obtidos em 2004 não podem ser completamente comparados aos resultados obtidos em 2003, visto que as duas edições foram desenhadas de forma diferente, no entanto, as precisão das respostas a perguntas especificas como as que têm como tipo de resposta um local ou um tempo, foi bastante alta em todas as sete línguas alvo. A introdução de perguntas do tipo "Como ...?" dificultou bastante a tarefa, tornando também difícil a avaliação das respostas a este tipo de perguntas.

Na *Apêndice A* estão descritos alguns dos sistemas participantes no QA@CLEF-2004.

Capítulo 3

O nosso sistema

3.1 Introdução

Neste capítulo é feita uma descrição da abordagem feita no Departamento de Informática da Universidade de Évora para um sistema de Perguntas e Respostas para a língua Portuguesa[12].

Este sistema de perguntas e respostas desenvolvido pelo Departamento de Informática da Universidade de Évora, quando começou a ser desenvolvido, começou como sendo um conjunto de módulos e ferramentas completamente separadas umas das outras.

Quando em 2004, o CLEF[1] adicionou a língua Portuguesa à sua tarefa Question Answer como sendo uma das possíveis línguas, foi aproveitada a oportunidade de participar com um sistema de perguntas e respostas, decidindo-se então juntar algumas das ferramentas linguísticas que já existiam e que estavam a ser desenvolvidas para criar um sistema de perguntas e respostas.

Este sistema de perguntas e respostas foi feito para a Língua Portuguesa sobre um domínio aberto, passando a ser um dos primeiro sistemas de perguntas e respostas desenhado exclusivamente para a Língua Portuguesa.

Este sistema tinha como principal objectivo conseguir responder a algumas perguntas feitas em Língua Portuguesa fazendo pesquisas em simples documentos de texto também em Língua Portuguesa.

O objectivo inicial da implementação deste sistema de perguntas e respostas não era implementar um sistema que conseguisse responder a todas as perguntas feitas em Língua Portuguesa mas apenas que conseguisse responder a algumas perguntas para que se conseguisse perceber se era possível implementar um sistema

[1] Cross Language Evaluation Forum

relativamente eficaz e quais os possíveis problemas que o sistema iria ter bem como as suas virtudes.

Os problemas que surgissem na implementação deste sistema iriam ajudar na implementação de um futuro sistema de perguntas e respostas que conseguisse obter resultados bastantes mais eficazes corrigindo-se os principais erros do sistema.

Este sistema de perguntas e respostas pode ser considerado como um protótipo de um futuro sistema de perguntas e respostas para a Língua Portuguesa podendo vir a tornar-se numa ferramenta bastante útil.

Este sistema tenta responder a perguntas em Língua Portuguesa fazendo as pesquisas em documentos de texto também em Língua Portuguesa. Um sistema deste tipo é considerado um sistema mono-lingue, respondendo apenas em perguntas em Língua Portuguesa sobre documentos da mesma língua.

Este sistema, está dividido por vários módulos interligados entre si. Os resultados destes módulos vão sendo passados de uns para os outros com o objectivo de ser chegar a uma resposta a partir da pergunta.

A maior parte dos os módulos usados neste sistema apenas têm significado quando estão interligados com os outros módulos do sistema, verificando-se assim a importância da interligação entre os módulos.

Este sistema de perguntas e respostas tem duas etapas importantes:

- Para cada pergunta é recuperado um conjunto de possíveis documentos relevantes
- Para cada um destes documentos relevantes que foram analisados anteriormente, é feita uma extracção dos factos que possam ser úteis através da interpretação das perguntas sobre a base de conhecimento que representa cada um dos textos. Quando uma resposta é encontrada o sistema mostra a resposta e identifica qual o documento que suporta a resposta

Este sistema de perguntas e respostas necessita de um módulo de recuperação de informação para definir um conjunto de possíveis documentos relevantes. Este módulo foi adicionado ao sistema devido a problemas de complexidade computacional.

A principal característica deste sistema é fazer analises profundas a um conjunto de documentos e obter uma representação semântica do seu conteúdo. Nesta abordagem, cada pergunta é também transformada numa representação semântica e depois um processo de inferência tenta obter uma resposta para a pergunta. Esta abordagem é bastante interessante, no entanto mostrou que tem graves problemas

de escalabilidade devido ao grande número de documentos disponível nas colecções de documentos.

Como já foi referido anteriormente, este sistema de perguntas e está dividido em vários módulos interligados entre si. Os principais módulos deste sistema são: o módulo de analise e processamento das perguntas, o módulo da geração de interrogações, o módulo de recuperação de informação, o módulo de interpretação pragmática e semântica, e o módulo de extracção da resposta.

O primeiro módulo recebe do utilizador uma pergunta em Língua Natural e faz uma análise da pergunta. Depois da pergunta estar analisada, o resultado é passado a um módulo de geração de interrogações, que por sua vez vão ser usadas num motor de busca para fazer a recuperação de documentos que possam ser relevantes para a pergunta. Depois dos possíveis documentos relevantes estarem recuperados, o módulo de interpretação semântica/pragmática gera um conjunto de bases de conhecimento que representam os documentos recuperados. Por fim, um último módulo faz a extracção das respostas a partir dos documentos recuperados pelo motor de busca.

Este sistema, para além de todos estes módulos referidos anteriormente tem um outro módulo que faz alguns pré-processamentos às perguntas e aos documentos de texto. Este módulo é executado antes de qualquer um dos outros módulos, produzindo resultados que vão ser analisados pelos outros módulos.

As perguntas necessitam de um pré-processamento para que consigam ser interpretadas de uma forma mais correcta, pois o módulo do tratamento e processamento das perguntas produz melhores resultados quando as perguntas estão em certos formatos ou quando algumas palavras das perguntas são removidas das perguntas.

Já os documentos de texto precisam de ser ser pré-processados para se fazer a sua tradução para um outro formato e para fazer a sua indexação.

A tradução dos documentos para para outro formato é necessária pois existem módulos que necessitam dos documentos de texto neste formato para conseguirem fazer o seu processamento.

A indexação dos documentos é feita para que o motor de busca consiga fazer a sua pesquisa no colecção de documentos disponíveis e possa obter os documentos relevantes para a pergunta.

Um dos requisitos deste sistema é a necessidade da existência de uma base de conhecimento de factos inferidos a partir dum conjunto dos documentos e duma ontologia com os conceitos que estão nos documentos

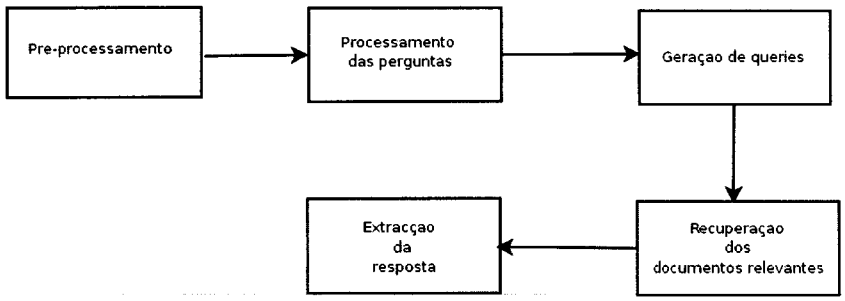

Figura 3.1: Arquitectura global do sistema.

A característica deste sistema é o uso de análises e representações semânticas em quase todo o processo de responder às perguntas, sendo aqui que este sistema se diferencia mais dos outros. Esta característica e contra-se principalmente na fase de processamento das perguntas e no pré-processamento dos documentos onde é usada esta representação semântica.

No processamento das perguntas e no pré-processamento dos documentos são criadas representações semânticas das perguntas e de cada um dos documentos, usando para isso ferramentas feitas com o objectivo de criar representações semânticas para textos e perguntas em Língua Portuguesa. A representação semântica escolhida para as perguntas foram as DRSs[2].

Esta representação semântica é também usada no módulo de extracção das respostas, para isso à feita uma unificação entre a representação semântica da pergunta e a representação semântica dos documentos relevantes, encontrando-se assim uma resposta para a pergunta.

3.2 Descrição do Sistema

A arquitectura global deste sistema é bastante simples e é composto por: um módulo de pré-processamento das perguntas e dos documentos de texto onde é feita a representação semântica dos textos e a indexação dos documentos, um módulo que faz a representação semântica das perguntas, um módulo de geração de interrogações que vão ser executadas no motor de busca a fim de obter os documentos relevantes e, um módulo de interpretação semântico/pragmática que dá origem a uma base de conhecimento e um módulo que faz a extracção das respostas.

[2]Discourse Representation Structure[11]

Como se pode ver na figura 3.1, a arquitectura global deste sistema é bastante parecida com muitos outros sistemas de perguntas e respostas como os que já foram descritos no capitulo 2.

A arquitectura global deste sistema tem um desenho bastante simples, no entanto, quando se olha para toda a arquitectura do sistema, esta pode tornar-se bastante mais complicada devido às várias ligações que existem entre os vários módulos. Na figura 3.2 pode-se ver mais em detalhe a arquitectura deste sistema de perguntas e respostas.

Este sistema começa por receber do utilizador uma pergunta em Língua Portuguesa. Depois do sistema ter recebido a pergunta em Língua , o módulo de pré-processamento entra em acção fazendo um pré-processamento das perguntas, alterando o seu formato, criando as representações semânticas dos documentos de texto e fazendo a sua indexação.

Depois das perguntas estarem pré-processadas são passadas ao módulo da geração de interrogações que faz a geração das interrogações a partir da representação semântica das perguntas. Depois das interrogações estarem feitas, o módulo de recuperação de documentos faz uma pesquisa sobre os documentos indexados como o objectivo de recuperar documentos relevantes para a pergunta.

Depois dos documentos relevantes terem sido recuperados é criada uma base de conhecimento ou uma colecção de factos a partir destes documentos, sendo criada uma base de conhecimento para cada um dos documentos.

Depois de estar criada uma base de conhecimento para cada um dos possíveis documentos relevantes, estas são passadas ao módulo de extracção de respostas junto com a representação semântica das perguntas e com a ontologia para se extrair a resposta.

O módulo de geração das DRSs dos documentos e das perguntas é o mesmo, estando este dividido em três sub-módulos. O primeiro destes três módulos cria uma árvore sintáctica dos documentos ou das perguntas através do analisador sintáctico PALAVRAS. Esta árvore sintáctica, por vezes não é perfeita, chegando mesmo a ter alguns erros graves.

Para resolver este problema foi criado um outro sub-módulo que faz a correcção de alguns destes casos em que a árvore sintáctica gerada não era a mais correcta.

Depois da árvore sintáctica estar gerada, são então criadas as DRSs, gerando-se assim as representações semânticas das perguntas e dos documentos de texto.

Este sistema necessita de uma ontologia e de uma base de conhecimento para conseguir fazer a extracção das respostas, sendo as duas construídas a partir dos possíveis documentos recuperados.

- Ontologia - A partir dos resultados da geração das DRSs e dos conceitos da ontologia geral é criada uma nova ontologia que contém os conceitos referidos nos documentos recuperados[13, 14].

  Este processo mostrou ser bastante complicado devido ao grande número de conceitos que são referidos nos documentos e à dificuldade de encontrar relações entre os vários conceitos.

  A ontologia obtida por este processo foi criada no formato OWL [3] e na forma lógica de programação ISCO [15], que permite uma integração de mecanismos de inferência parecidos aos do Prolog, mas com classes, hereditariedade e algoritmos com *constraints*.

- Base de conhecimento - A partir desta ontologia e da representação semântica de cada uma das frases pode-se obter uma interpretação de cada frases que vai dar origem a um conjunto de factos que vão ser adicionados na base de conhecimento[16].

  No entanto, esta tarefa mostrou-se computacionalmente complexa, quer ao nível de tempo como ao nível de espaço, gerando bases de conhecimento muito grandes, criando assim muitos problemas para o processo de inferência da resposta.

  Para resolver este problema foi decidido diminuir o número de documentos relevantes para cada pergunta através de processos de *Intormation Retrieval*, sendo depois criada a base de conhecimento.

  A base de conhecimento geral foi criada com factos extraídos da colecção de documentos de regras e factos importados de outras aplicações.

3.2.1 Pré-processamento e indexação

O pré-processamento das perguntas é bastante simples, sendo apenas feita uma simples análise às perguntas em língua natural onde é alterado o formato de algumas perguntas para que o sistema consiga obter melhores resultados. A arquitectura de todo o pré-processamento está representado na figura 3.3

O pré-processamento dos documentos de texto tem duas partes, sendo primeiro feita uma análise semântica onde é criada uma representação semântica dos documentos e numa segunda parte é feita a indexação dos documentos de texto. A representação semântica das perguntas é baseada nas DRSs[4][11].

[3]Ontology Web Language

[4]Discourse Representation Structure[11]

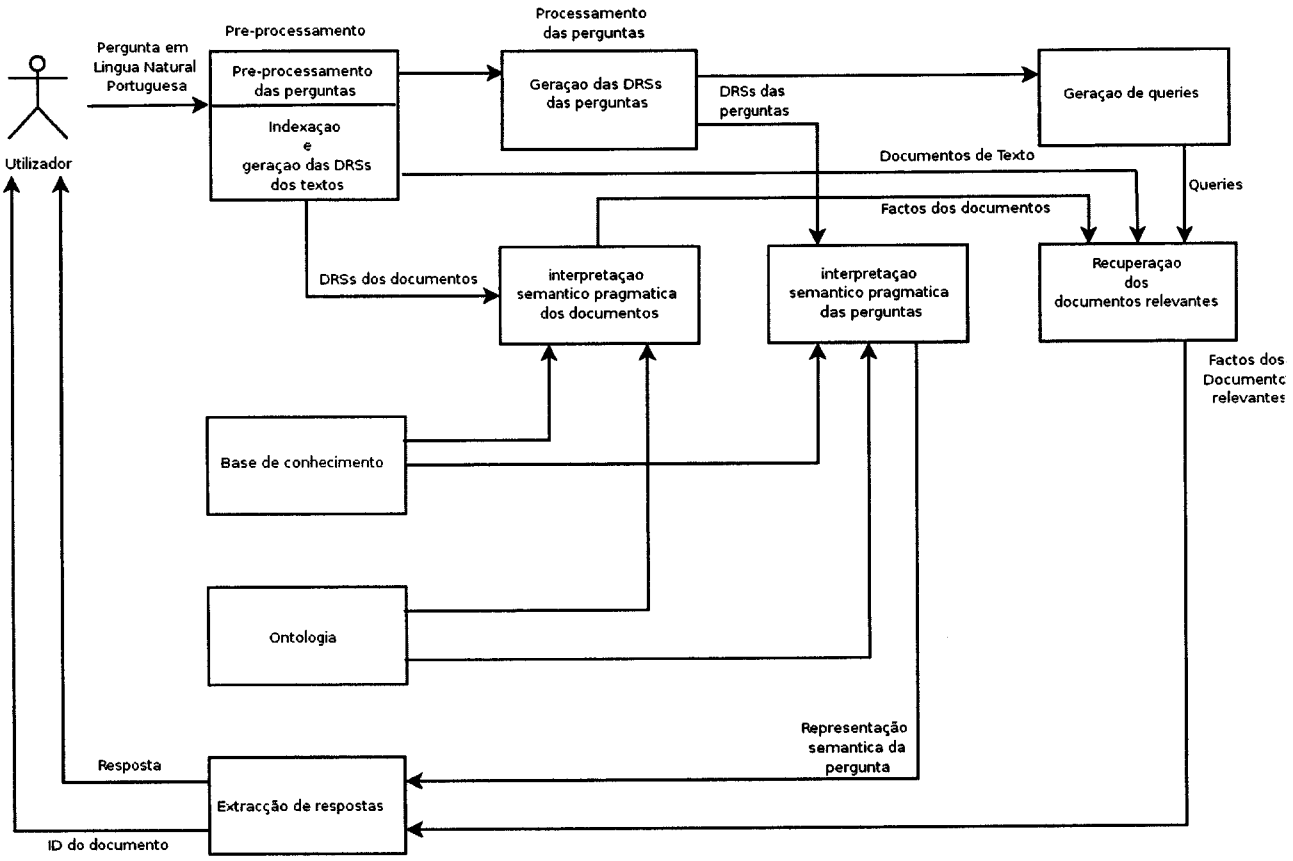

Figura 3.2: Pormenor dos vários módulos do sistema.

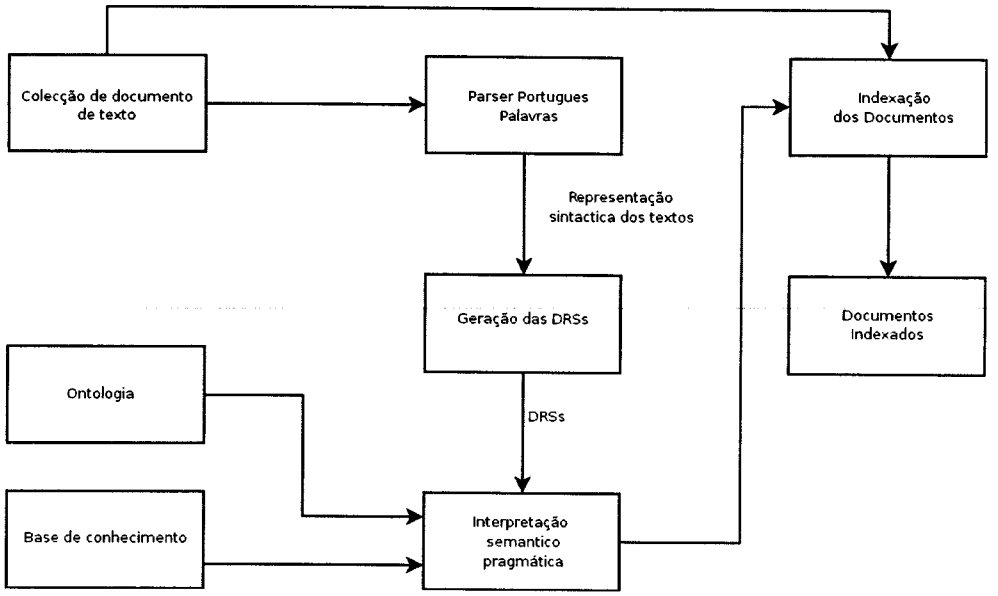

Figura 3.3: Pre-processamento do sistema.

O processo de geração da representação semântica dos documentos é feita a partir da representação sintáctica dos documentos de texto.

No pré-processamento dos documentos foi também necessário fazer a indexação dos documentos de texto por um motor de busca para que mais tarde, o módulo de recuperação de informação consiga obter os documentos relevantes para a pergunta.

Depois de todo o pré-processamento estar feito, tem-se as perguntas em língua natural possivelmente re-escritas noutra forma, uma representação semântica dos documentos de texto e a indexação dos documentos feita por um motor de busca.

3.2.2 Processamento das perguntas

Depois de todo o pré-processamento estar feito, a primeira tarefa a ser feita pelo sistema é uma análise semântica às perguntas em Língua Natural. Estas perguntas em Língua Natural podem já não estar como no seu formato original, podendo já terem sido re-escritas. Neste caso, se as perguntas já foram re-escritas a geração da representação semântica é criada a partir desta nova pergunta.

O processo de criar a representação das DRSs das perguntas está dividido em três módulos:

- representação sintáctica.

- correcção da representação sintáctica.
- geração da representação semântica a partir da representação sintáctica.

Este processo de análise e representação semântica usa as mesmas ferramentas e metodologias usadas no pré-processamento dos documentos onde foram geradas as suas representações semânticas.

Depois da representação semântica das perguntas estar completa é feita uma interpretação semântico/pragmática da pergunta tendo em conta a ontologia geral de conceitos e uma base de conhecimento com algum conhecimento geral.

3.2.3 Geração de interrogações

Depois das perguntas estarem processadas e terem a sua representação semântica feita, são criadas algumas interrogações a partir desta representação. Estas interrogações são geradas a partir da representação semântica das perguntas usando algumas das palavras chave da pergunta e ignorando outras palavras que não têm qualquer significado na interrogação que vai ser executada num motor de busca para obter os documentos que possam relevantes para a pergunta.

Neste módulo são criadas 3 interrogações para cada pergunta, sendo depois estas usadas no módulo de recuperação de informação para obter os documentos relevantes para a pergunta.

3.2.4 Recuperação de informação

O módulo de recuperação de informação obtém os documentos que possam ser relevantes para a pergunta através de técnicas e ferramentas tradicionais de recuperação de informação.

Usando um motor de busca e as interrogações que foram geradas para as perguntas junto com a colecção de documentos de texto indexados vão-se obter os documentos relevantes para a pergunta.

Neste sistema, a pesquisa dos documentos relevantes é feita sobre os documentos de texto indexados, no entanto os documentos relevantes que vão ser usados nos módulos seguintes são os documentos com a sua representação semântica.

3.2.5 Interpretação pragmático/semântica dos documentos relevantes

Para se obter um conjunto de factos para cada frase dos documentos relevantes é necessário usar a ontologia para se obter o significado de cada uma das frases dos documentos de texto relevantes.

Este módulo de interpretação semântico/pragmático recebe como *input* as representações semânticas dos documentos relevantes para construir uma base de conhecimento com base nos factos referidos nos documentos de texto.

3.2.6 Extracção das respostas

Depois ter sido feito a recuperação dos documentos relevantes para a pergunta, é altura de se tentar extrair uma resposta dos documentos.

Este módulo recebe um conjunto de bases de conhecimento junto com a representação semântica das perguntas obtida a partir da interpretação pragmático / semântica das perguntas.

Nesta fase, o sistema vai trabalhar com as bases de conhecimento criadas para cada documento relevante e com a representação semântica da pergunta.

Com estes dados o sistema tenta inferir uma resposta usando uma forma lógica de programação chamada ISCO[17, 15] que permite fazer uma integração de processos de inferência como no Prolog mas com o uso de classes, hereditariedade e algoritmos com *constraints*.

3.3 Descrição dos módulos

3.3.1 Pré-processamento e indexação

Alteração das perguntas

Este módulo recebe como entrada as perguntas em Língua Portuguesa e altera o formato de algumas perguntas para que o sistema consiga obter melhores resultados. Este processamento é feito através de simples técnicas de *pattern matching*.

Este simples módulo foi feito em *perl* devido às suas expressões regulares e à facilidade de trabalhar com elas.

As regras usadas para fazer este *pattern matching* foram feitas manualmente, usando para isso uma colecção de perguntas em Língua Portuguesa. Esta colecção de perguntas foi feita a partir de perguntas simples que são frequentemente feitas. Esta colecção de perguntas foi feita de forma a que fizesse uma representação das perguntas mais frequentes usadas no dia a dia.

Nesta colecção de perguntas foram também usadas perguntas traduzidas manualmente da colecção de perguntas do QA@CLEF-2003 em Espanhol e da colecção de perguntas usadas no TREC-QA em Inglês.

As colecções de perguntas do CLEF e do TREC foram usadas principalmente para se ter uma ideia de que tipo de perguntas é que poderiam surgir e ser feitas ao sistema num WORKSHOP semelhante. Com a colecção de perguntas feita,

as perguntas foram analisadas manualmente para fazer uma categorização das perguntas.

Depois de se saber qual o tipo de perguntas mais frequente na colecção de perguntas, estas foram testadas para saber qual iria ser o resultado da geração das DRSs. Nestes testes foram verificadas a árvore sintáctica e a DRS geradas para a pergunta.

No caso da arvore sintáctica da frase não ser a mais indicada para a pergunta, alterou-se o formato da frase manualmente de forma a que a árvore sintáctica da pergunta fosse mais correcta correcta. Depois de se ter encontrado uma nova forma para a pergunta em Língua Natural que fosse capaz de gerar uma árvore sintáctica correcta para a pergunta, foi criada uma expressão regular que conseguisse fazer as alterações necessárias a essa pergunta mas também a outras perguntas com do mesmo tipo, conseguindo-se assim fazer o tratamento de várias perguntas ao mesmo tempo apenas com uma expressão regular.

No caso da DRS da pergunta não estar correcta, tenta-se também alterar a forma como as perguntas são feitas de forma a que possam ser geradas DRSs correctas para as perguntas, sendo também criadas expressões regulares que consigam alterar a forma como a pergunta é feita.

Com a alteração da forma como algumas perguntas são feitas consegue-se que sejam produzidas arvores sintácticas mais correctas gerando-se assim melhores DRSs para as perguntas. Nos exemplos seguintes pode-se ver algumas destas perguntas junto com a sua nova versão:

```
``A que velocidade viaja a luz?'' -> ``que velocidade viaja a luz?''
``Mencione um bonecreiro.''       -> ``que um bonecreiro.''
```

As perguntas são alteradas são perguntas que se encontram dentro de um certo padrão e bastante frequentes, sendo alteradas a forma das perguntas que se encontrem nos seguintes formatos, usando para isso as regras indicadas para se alterar a pergunta:

- Se for encontrada a palavra "local" esta é trocada por "lugar".
- A palavra "onde" é trocada pela a expressão "em que local".
- Se aparecer uma pergunta do tipo "De que cor é a neve?" então a pergunta é transformada em "Qual a cor de neve?"
- Se na pergunta aparecerem as palavras "nomei", "indique" ou "mencione" então são substituídas pela palavra "que"
- Se na frase aparecer a expressão "a que" ou "o que" são substituídas pela palavra "que".

- Se aparecer uma pergunta do tipo "Em que cidade se encontra a prisão de San Vittore" é trocada a expressão "Em que" pela expressão "Qual a".

- Se a pergunta for do tipo "Como morreu Jimi Hendrix?", então a pergunta é alterada para a forma "morreu Jimi Hendrix de quê?".

Como se pode ver nos exemplos e nas regras anteriores, são tratadas bastantes perguntas com poucas regras. Com estas regras consegue-se alterar a forma como são feitas muitas perguntas pois são regras muito genéricas. Com o uso destas regras consegue-se assim que muitas perguntas que iriam gerar DRSs incorrectas sejam alteradas de forma a gerarem DRSs mais correctas de forma a representarem melhor a pergunta.

Indexação dos documentos

A indexação dos documentos de texto é uma das fases mais importantes do sistema, pois, sem existir uma indexação não é possível fazer a recuperação dos documentos relevantes para a pergunta. A indexação dos documentos foi feita pelo motor de busca que mais tarde faz a recuperação dos documentos. O motor de busca usado foi o SINO[5] [18, 19] criado no Australasion Legal Information Institute. O SINO permite a criação de ficheiros de indexação invertidos e usa informação especifica para língua Portuguesa, usando *stop words*, *lemmatization*. Este motor de busca foi alterado para para usar um conjunto de *stop words* portuguesas como artigos, pronomes e preposições e para transformar cada palavra na sua forma básica, usando para isso o léxico português POLARIS.

Para fazer a indexação dos documentos de texto com SINO foi preciso fazer algumas alterações aos ficheiros de texto para que o sino pudesse fazer a sua indexação da forma mais correcta. Este motor de busca permite fazer pesquisas em documentos de texto, fazendo a busca em todo o documento, ou fazendo a busca apenas em algumas secções do documento. Para tirar partido desta funcionalidade do SINO foi necessário acrescentar aos documentos de texto secções para que se pudesse fazer as pesquisas nessas secções dos documentos de texto.

Para construir estas secções nos documentos de texto é apenas necessário acrescentar algumas linhas ao ficheiro de texto. Estas linhas, são linhas especiais com uma sintaxe que o SINO percebe e que limitam as várias secções do ficheiro de texto.

Estes limitadores das secções dos ficheiros de texto são bastante parecidos com etiquetas de XML, sendo apenas necessário indicar onde começa uma nova secção com a linha ``<!-- sino section SECÇÃO -->'' onde "SECÇÃO" é o nome da

[5]Yet another search engine for the Web

secção, e para terminar a secção de um ficheiro de texto usa-se a linha ``<!-- end section -- > ''.

Este processamento aos documentos de texto foi um processamento simples, usando para isso algumas ferramentas escritas em "Perl" para automatizar o processo de adicionar estas *tags* nos ficheiros para que se pudesse aproveitar todas as funcionalidades que o motor de busca possui.

Representação semântica dos documentos

No pré-processamento dos documentos de texto é também necessário fazer uma representação semântica dos documentos. Esta representação dos documentos está dividida em três fases. Numa primeira fase, é feita uma analise sintáctica junto com a etiquetação morfosintáctica dos documentos de texto. Nesta representação sintáctica as palavras são reduzidas à sua forma mais básica, existindo nesta representação várias informações disponíveis, como o tipo da palavra e a sua posição no texto.

Esta representação é feita através de um analisador sintáctico para a língua Portuguesa chamado PALAVRAS[20] desenvolvido o VISL[6] que é um projecto de investigação e desenvolvimento do *Institute of Language and Communication* da *University of Southern Denmark*.Esta ferramenta, embora seja um analisador e interpretador sintáctico, usa também alguma informação semântica, construindo assim melhores árvores sintácticas para os textos.

Esta representação sintáctica é depois transformada em predicados Prolog, que representam a árvore sintáctica do documento de texto através do uso de predicados Prolog.

Depois da árvore sintáctica estar gerada em Prolog, é então feita a representação semântica do documento de texto. A representação semântica dos documentos de texto é feita sobre árvore sintáctica dos documentos em Prolog.

A análise semântica dos documentos de texto é feita em Prolog, uma vez que a árvore sintáctica já está transformada num conjunto de predicados Prolog. Esta análise é feita usando pequenas árvores sintácticas às quais estão associadas algumas informações semânticas, sendo a arvore sintáctica analisada aos poucos até se fazer toda a representação semântica da arvore sintáctica do documento.

Estas pequenas árvores sintácticas vão sendo unificadas recursivamente com as várias árvores sintácticas do documento de texto com a ajuda algumas regras de gramática.

[6]Visual Interactive Syntax Learning

A representação semântica do documento é feita usando as DRSs que são uma boa forma de fazer representações semânticas.

As DRS assentam na teoria DRT que é uma das muitas teorias de semântica dinâmica que têm vindo a aparecer ao longo dos últimos 20 anos. Esta teoria envolve dois passos: primeiro é construída a representação semântica(DRS's), e depois é construído um modelo de interpretação sobre esses DRS's. As DRS's são constituídas por duas partes:

1. Referentes do discurso.

2. Por um conjunto de condições.

As DRS's são um conjunto de referentes e de condições, onde o conjunto de referentes da DRS pode ser considerado como o "mundo" da DRS e as condições podem ser consideradas as restrições que existem no mundo.

No seguinte exemplo a árvore sintáctica gerada para a "O gato do João comeu o rato do Manuel".

```
STA:fcl
=SUBJ:np
==>N:art('o' M S <artd>)          O
==H:n('gato' M S <H>)    gato
=P:v-fin('comer' PS 3S IND)       comeu
=ACC:np
==>N:art('o' M S <artd>)          o
==H:n('rato' M S <H> <Adom>)      rato
==N<:pp
===H:prp('de' <sam->)    de
===P<:np
====>N:art('o' M S <artd> <-sam>)          o
====H:prop('Manuel' M S)          Manuel
=.
```

Como se pode ver na representação sintáctica da frase "O gato do João comeu o rato do Manuel" produzida pelo VISL, cada palavra é reduzida à sua forma mais reduzida, havendo indicação do tipo de cada palavra junto com mais algumas informação que podem ser bastantes úteis, como o género e o número das palavras, o tipo de cada verbo a a sua conjunção. Esta árvore sintáctica é bastante simples,

sendo facilmente interpretada e compreendida por um ser humano que consegue saber facilmente relações pai/filho que existe entre cada parte da árvore. No entanto torna-se bastante complicado fazer análises sobre esta árvore. Para resolver este problema foi usado um pequeno programa que converte esta árvore sintáctica produzida pelo VISL em predicados Prolog que geram uma árvore sintáctica igual a esta mas em Prolog.

```
sentence(syn(
     sta(fcl,
         subj(np,
                      n(art('o','M','S','<artd>'),'O'),
           h(n('gato','M','S','<H>'),'gato')
                     ),
                p(v_fin('comer','PS','3S','IND'),'comeu'),
                acc(np,
                    n(art('o','M','S','<artd>'),'o'),
                    h(n('rato','M','S','<H>','<Adom>'),'rato'),
                    n(pp,
                      h(prp('de','<sam->'),'de'),
                      p(np,
                        n(art('o','M','S','<artd>','<-sam>'),'o'),
                        h(prop('Manuel','M','S'),'Manuel', '.')
                       )
                      )
                    )
               )
          )
        ).
```

Na representação anterior encontra-se um predicado Prolog da árvore sintáctica produzida pelo VISL para a mesma frase. Como se pode ver esta árvore é um simples predicado Prolog que representa uma simples árvore. Usando este predicado Prolog em vez da representação sintáctica gerada pelo VISL torna-se bastante mais simples fazer análises sobre a árvore sintáctica. Este predicado mantém toda a informação que existe na representação gerada pelo VISL, estando assim toda a informação disponível que existia na representação produzida pelo VISL.

Sobre esta representação da árvore sintáctica da frase "O gato do João comeu o rato do Manuel" vai ser criada a DRS para a frase, sendo gerada a seguinte DRS:

```
drs(
    [def-A-m-s,def-B-m-s,def-C-m-s],
    [gato(A),comer(A,B),rato(B),rel(de,B,C),nome(C,'Manuel')]
  )
```

Como se pode ver no exemplo anterior, a DRS para a frase "O gato do João comeu o rato do Manuel." é constituída por duas partes: os referentes do discurso e as condições da DRS. Esta DRS tem 4 referentes, um para cada entidade da frase que possa pertencer ao *mundo* da DRS e 7 condições. Nas condições desta DRS pode-se ver que o referente *A* é um gato que o referente *B* se refere a um nome("João"), que *C* é um rato e *D* também é um nome("Manuel"). Nas condições da DRS existem havendo várias relações entre cada os referentes, gerando-se assim relações e restrições sobre os referentes do discurso. Com as relações que existem entre cada um dos referentes sabe-se que existe uma ligação entre "gato" e "João" e que a palavra que faz esta relação é a palavra "de", existe uma outra relação entre "gato" e "rato" e relação entre estes dois referentes é o verbo "comer". Por fim existe uma outra relação entre "rato" e "Manuel" sendo aqui a palavra "de" que faz a ligação entre "rato" e "Manuel".

Nesta representação semântica os referentes do discurso têm disponível o seu género e número. Para alguns referentes existe também o tipo do referente. No caso desta DRS, pode-se ver também que todos os referentes são definidos, masculinos e singulares.

Os referentes estão representados num formato que permite saber qual o tipo de referente, o *nome* do referente o género e o número. Existem dois tipos de referentes, os que têm a informação do tipo do seu tipo e os que não têm esse tipo de informação. Os referentes que não têm a informação em relação ao tipo do referente estão representados na forma *tipo(palavra)-NOME-género-número*, onde *tipo* representa o tipo do referente, *palavra* representa a palavra que está associada ao referente *NOME* é o nome do referente, *género* é o género do referente e *número* é o número do referente. O género do referente é representado por *'f'* ou *'m'* no caso de ser respectivamente feminino ou masculino, ou então *m/f* no caso em que o referente não tem um género definido. No caso do número do referente, este pode ser plural ou singular, sendo este representado por *'p'* ou *'s'*.

No caso de ser um referente que não tem qualquer informação em relação ao seu tipo, este é representado na seguinte forma *DEF-NOME-género-número*. Neste tipo de referentes o que aparece de novo é o *DEF*, que simplesmente indica se o referente está ou não definido, sendo representado como *'def'* no caso de ser definido ou *'indef'* no caso de ser indefinido. Neste tipo de referentes, o nome, o género e o número dos referentes são representados exactamente da mesma forma que no outro tipo de referentes e têm exactamente o mesmo significado.

3.3.2 Processamento das perguntas

Neste módulo vai ser criada uma representação semântica das perguntas bem como a sua interpretação semântico/pragmática.

Esta representação semântica das perguntas é feita exactamente da mesma forma que a representação semântica dos documentos de texto usando as mesmas ferramentas e as mesmas metodologias usadas na análise semântica das perguntas, obtendo assim resultados equivalentes para as perguntas.

Na geração das DRS's das perguntas são acrescentadas algumas funcionalidades, havendo um novo tipo de referentes para os referentes relacionados com as interrogativas como por exemplo "Quem", "Como" e "Qual". Estes referentes têm associados o seu tipo e a palavra que deu origem ao referente. Os referentes *interrogativos* são representados na seguinte forma: *interr(PALAVRA)-NOME-GÉNERO-NÚMERO.*

O que aparece de novo nesta representação é *interr(PALAVRA)* que indica que o referente está associado a uma interrogativa e foi gerado pela palavra *PALAVRA.*

O processamento e geração das DRSs para as pergunta é feita da mesma forma que é feito para os documentos de texto. Para a pergunta "Quem matou uma gata do Manuel?" iria ser gerada a seguinte DRS:

```
drs(
    [interr(quem)-A-'m/f'-s,indef-B-f-s,def-C-m-s],
    [matar(A,B),gata(B),rel(de,B,C),nome(C,'Manuel')]
)
```

Nesta DRS, pode-se ver que no referente da interrogativa está associada a palavra *'quem'*, que o seu género não está definido e que é singular. Pode-se também ver que o referente *B* é um referente *indefinido* isto por causa de ser *'uma gata'* e não *'a gata'*.

Na geração desta DRS não são mostrados todos os passos intermédios, no entanto, a geração da DRS passa por todos os processos mostrados anteriormente, sendo gerada a sua arvore sintáctica em VISL, sendo depois convertida para predicados Prolog, feita a correcção da arvore sintáctica e por fim é gerada a DRS final.

O módulo da geração das DRSs, tanto para as perguntas como para os documentos, é composto por três sub-módulos, sendo a sua arquitectura representada na figura 3.4

Depois da representação semântica das perguntas estar completa é feita uma interpretação semântico/pragmática da pergunta tendo em conta a ontologia geral de conceitos e uma base de conhecimento com algum conhecimento geral. Esta

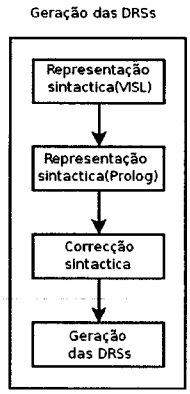

Figura 3.4: Arquitectura do módulo da geração das DRSs

interpretação semântico/pragmática vai dar origem à versão final da representação semântica.

Para a pergunta "Quem matou o gato do Manuel" iria ser gerada a seguinte DRS:

```
drs(
     [interr(quem)-A-'m/f'-s,indef-B-m-s,def-C-m-s],
     [matar(A,B),
      gato(B),rel(de,B,C),
      nome(C,'Manuel')]
)
```

Ao ser feita a interpretação semântico/pragmática desta DRS ir-se-ia obter a seguinte representação semântica:

```
drs(
     [interr(quem)-A-'m/f'-s,indef-B-m-s,def-C-m-s],
     [matar(A,B),
      gato(B),owns(C,B), person(C),
      nome(C,'Manuel')]
)
```

Esta ultima representação semântica é a que vai ser avaliada nas bases de conhecimento que foram obtidas a partir dos possíveis documentos relevantes obtidos pelo sistema de recuperação de informação.

3.3.3 Geração de interrogações

Depois de todo o processamento das perguntas estar feito e geradas as DRSs das perguntas, vão ser geradas interrogações para as perguntas. Estas interrogações depois vão depois ser executadas no motor de busca para obter os documentos relevantes para a pergunta.

Estas interrogações são criadas a partir das DRSs geradas para as perguntas, usando para isso os referentes de discurso junto com as condições. Apenas algumas palavras da pergunta em Língua Natural é que passam para a DRS, sendo estas as palavras mais relevantes para a pergunta, enquanto que outras palavras que não têm muito interesse para a pergunta são ignoradas.

Estas interrogações são bastantes relevantes, pois as palavras pouco relevantes para a pergunta já foram ignoradas.

As interrogações que são geradas são interrogações booleanas baseadas nas condições das DRSs das perguntas. Estas interrogações usam operadores lógicos entre as várias palavras para gerar uma interrogação booleana que representa a pergunta. Os operadores booleanos que são usados são o *AND* e o *OR*. Nestas interrogações são usadas várias combinações destes operadores. Os operadores usados na construção das interrogações dependem do tipo de cada um dos referentes e da especificidade que se pretende dar à interrogação.

São geradas 3 interrogações para cada pergunta, sendo gerada uma interrogação muito especifica que irá obter poucos resultados mas muito precisos, uma interrogação muito geral que obtém muitos resultados mas pouco precisos e uma outra interrogação que não é muito específica nem é muito geral.

São geradas 3 interrogações para cada pergunta para que se consiga obter um número de documentos relevantes suficiente para que o sistema consiga obter uma resposta.

Se apenas fosse usada a interrogação muito específica corria-se o risco de se encontrarem poucos documentos relevantes ou até de não se encontrarem documentos relevantes para a pergunta, se fosse usada a interrogação mais geral, ir-se-ia encontrar imensos documentos mas muito pouco relevantes.

Para resolver este problema são então usadas estas três interrogações para se chegar a um certo número de documentos relevantes, começando-se por obter os documentos relevantes através da interrogação mais especifica, passando pela intermédia até chegar à interrogação mais geral.

Existem algumas regras simples que ajudam na construção das interrogações para as perguntas. Com estas regras consegues-se ignorar algumas palavras e alguns tipos de referentes como alguns verbos e artigos que não devem estar na

interrogação, pois de outra forma iria ser gerada uma interrogação que quando fosse executada no motor de busca iria obter documentos que não eram relevantes para a pergunta.

As palavras e os tipos de referentes que são ignorados na geração das interrogações são verbos como *haver*, *nomear*, *mencionar*, *indicar*, *intitular*, *ficar*, *significar*, *ter* e *ser*.

Estes verbos, embora possam ser bastante importantes no contexto da pergunta não podem ser incluídos na interrogação pois iriam ser encontrados documentos pouco relevantes para a pergunta.

Na pergunta "Mencione o nome de um poema." ver que a palavra "Mencione" é bastante importante para a pergunta, mas no entanto se fosse construída uma interrogação com a palavra "Mencione", os documentos recuperados não iriam ser muito relevantes pois iriam ser recuperados documentos com a palavra "Mencione". Estes documentos não teriam qualquer interesse para a resposta à pergunta, sendo assim ignorados estes tipos de palavras para que não aconteçam casos parecidos com este.

Para além destas palavras existem outras que são ignoradas, como por exemplo *ele*, *ela* e outros pronomes. Embora existam vários tipos de palavras que são ignorados na construção da interrogação as mais importantes são os verbos mencionados em cima.

Os nomes próprios, quando têm mais do que um nome, na DRS são representados no seguinte formato:*NOME1_NOME2_...*.

- João Manuel → João_Manuel
- Presidente da República → Presidente_da_Republica

Neste último exemplo pode-se ver a representação de alguns nomes na DRS, verificando-se este formato para todos os nomes próprios ou palavras capitalizadas.

Como estes nomes são bastante importantes para a pergunta estes têm que pertencer à interrogação gerada para essa pergunta. O motor de busca não pode fazer a pesquisa usando estes nomes separados pelo '_' pois iria encontrar poucos ou nenhuns documentos. O motor de busca iria procurar documentos que contenham o texto "João_Manuel" em vez de "João Manuel", encontrando assim documentos que não eram nada relevantes para a pergunta.

Para resolver este problema foi então criada uma regra que remove os '_'s dos nomes de forma a que se obtenha um nome próprio correcto.

Uma das regras que foi imposta foi o uso dos mesmos termos nas três interrogações que são geradas. As três interrogação geradas para cada uma das pergunta

usam exactamente as mesmas palavras diferindo apenas as condições lógicas que existem entre elas.

Na interrogação mais geral é feito um *AND* entre todas as palavras da interrogação, obtendo-se assim uma interrogação do tipo:

- SEARCH text(TERMO1 AND TERMO2 AND ...).

Esta interrogação vai ser a mais especifica, obtendo-se assim os resultados mais relevantes para a pergunta.

A interrogação mais geral é idêntica à mais especifica, substituindo-se os *AND* por *OR*. Esta interrogação é uma disjunção de todos os termos em vez de uma conjunção, ficando com o seguinte formato:

- SEARCH text(TERMO1 OR TERMO2 OR TERMO3 OR ...).

Neste tipo de interrogações é necessário ter alguns cuidados especiais em relação aos nomes próprios e aos nomes capitalizados. Uma vez que os nomes próprios são representados na DRS como: "João_Manuel", e estes por sua vez são transformados em "João Manuel" não podem entrar directamente para uma interrogação deste tipo e que é composta por uma disjunção de todos os termos, pois iria-se obter uma interrogação do tipo *João OR Manuel*. Esta interrogação não é o que se pretende, pois assim estaria-se à procura de documentos contenham o nome João ou o nome Manuel, enquanto que o que se pretende são documentos que contenham o nome João e Manuel.

Para resolver este problema criou-se uma outra regra que transforma as várias palavras de cada nome próprio com mais do que um nome ou as várias palavras de um conjunto de palavras capitalizada numa conjunção das várias palavras, sendo esta conjunção considerada como um único termo da interrogação.

- João Manuel → João_Manuel → (João AND Manuel)
- Presidente da República → Presidente_da_Republica → (Presidente AND da AND Republica)

Neste exemplo pode-se ver as várias transformações que estes tipos de nomes tiveram ao longo da geração das interrogações, retirando-se primeiro os "_" e substituindo-se depois com *ANDs*. Com estas regras é transformado um nome de várias palavras numa conjunção das várias palavras dos nomes.

Esta interrogação é a mais geral e a que obtém mais resultados quando é executada, no entanto os resultados obtidos pela execução desta interrogação são muito pouco relevantes para a pergunta pois a interrogação é muito geral.

A interrogação intermédia em termos de especificidade é feita através de conjunções e disjunções, sendo uma conjunção entre a palavra associada á primeira condição da DRS que é considerada como sendo a *head* da pergunta e a disjunção de todas as outras termos. Este tipo de interrogação fica representada no seguinte formato:

- SEARCH text(TERMO1 AND (TERMO2 OR TERMO3 OR ...))

Neste tipo de interrogação, se aparecer um nome próprio com mais do que um nome ou um conjunto de palavras capitalizadas, então este é tratado e representado como na interrogação mais geral, sendo feito um *AND* de todos os termos do nome próprio ou do conjunto de palavras, sendo esta conjunção considerada com um único termo da interrogação.

De seguida pode-se ver o exemplo das interrogações geradas para a pergunta "Em que cidade se encontra a prisão de San Vittore?":

```
SEARCH text(cidade AND encontrar AND prisão  AND (San  AND Vittore))

SEARCH text(cidade AND (encontrar OR prisão OR (San AND Vittore)))
```

Neste caso foram apenas geradas duas interrogações, pois com estas duas interrogações o motor de busca conseguiria obter 50 documentos relevantes.

No caso de não se conseguirem obter 50 documentos relevantes com a execução destas duas interrogações iria então ser gerada a terceira interrogação que seria bastante mais geral. Essa interrogação mais geral para a mesma pergunta seria:

```
SEARCH TEXT(cidade OR encontrar OR prisão OR (San  AND Vittore))
```

Como se pode ver, esta ultima interrogação é bastante mais geral, sendo uma disjunção das várias palavras da interrogação. Com esta interrogação o sistema consegue obter os 50 documentos pois é uma interrogação muito geral.

Com a execução da interrogação mais geral conseguem-se obter bastantes documentos relevantes, no entanto estes documentos são muito pouco relevantes pois basta que apareça uma das palavras da interrogação no documento para que o documento seja recuperado.

3.3.4 Recuperação de informação

Este módulo tem como objectivo recuperar todos os documentos que possam ser relevantes para a pergunta. Depois de geradas todas as interrogações para as perguntas estas são executadas neste módulo. Neste módulo as interrogações de cada pergunta são executadas no motor de busca para se encontrarem os documentos relevantes.

Este módulo de recuperação de informação pode ser considerado como um módulo opcional, no entanto não deixa de ser bastante importante no sistema de perguntas e respostas.

Este módulo serve para identificar os documentos que possam ser relevantes para uma pergunta, sendo depois a resposta procurada neste conjunto de documentos.

A importância da identificação dos documentos relevantes deve-se á imensa colecção de documentos de texto, sendo uma colecção de grandes dimensões e que pode crescer a qualquer altura. Tendo em conta esta característica da colecção de documentos torna-se quase impossível analisar todos os documentos para se encontrar uma resposta para a pergunta, visto que o processamento dos documentos para encontrar uma resposta torna-se bastante *pesado* em termos de tempo e de espaço principalmente na criação das bases de conhecimento.

Se o sistema procura-se as resposta em todos os documentos da colecção, o sistema iria tornar-se bastante lento e ocupando muitos recursos computacionais. Assim, para resolver este problema, usa-se o motor de busca para recuperar os documentos que possam ser relevantes , não sendo assim necessário procurar as respostas em todos os documentos. Desta forma o sistema pode procurar as respostas apenas nos documentos que são identificados pelo motor de busca, tornando-se mais fácil encontrar uma resposta para a pergunta.

O motor de busca que foi usado para fazer a recuperação dos documentos relevantes foi o Sino[7], feito pela equipa de Andrew Mowbray no Australasian Legal Information Institute, é um motor de busca com licença livre e desenhado para trabalhar em conjunto com servidores de *http* ou com outras aplicações. Este motor de busca, tal como muitos outros, necessita que a colecção de documentos onde a pesquisa vai ser feita já tenham sido indexada. Para fazer as pesquisas de documentos este motor de busca usa interrogações booleanas.

A indexação dos documentos bem como a geração das interrogações para as perguntas já foram feitas em módulos anteriores, estando disponível tudo o que é necessário para que o motor de busca possa recuperar os documentos relevantes.

[7]Yet another search engine for the Web

Este motor de busca recebe as interrogações que foram geradas anteriormente pelo módulo de geração de interrogações para recuperar os documentos que são relevantes para a pergunta. São geradas 3 interrogações para cada pergunta, mas no entanto apenas são usadas as interrogação que forem necessárias para se obterem os 50 documentos relevantes.

Se o motor de busca obter 50 documentos relevantes com a primeira interrogação, então a pesquisa de documentos relevantes pára e não são executadas mais interrogações. No caso de não serem encontrados os 50 documentos relevantes com a execução da interrogação mais especifica vai ser usada a segunda interrogação mais especifica e assim sucessivamente até se obterem 50 documentos relevantes para a pergunta. O número de 50 documentos relevantes achou-se suficiente para encontrar algumas respostas, no entanto este valor pode ser alterado se não se encontrarem respostas neste conjunto de documentos relevantes. No entanto, verificou-se que normalmente quando não se encontram as respostas neste conjunto de documentos, dificilmente se irá encontrar uma resposta recuperando mais documentos, pois os documentos recuperados para além destes são muito pouco relevantes

O SINO ordena os resultados da pesquisa através de uma função interna que calcula a relevância para cada um dos documentos para a interrogação que foi executada, dando assim um peso maior aos documentos mais relevantes para a pergunta.

Com esta ordenação o sistema sabe quais são os documentos com maior relevância para a pergunta, sendo estes os primeiros a serem analisados. Estes documentos com maior relevância são os que têm uma maior probabilidade de terem uma resposta para a pergunta. Embora estes documentos são os que têm uma relevância maior para a pergunta não há quaisquer garantias de que possam ter uma resposta para a pergunta e mesmo que tenham uma resposta para a pergunta não há garantias de que o sistema a consiga extrair.

No caso de terem sido usadas mais do que uma interrogação para fazer a recuperação dos documentos, os resultados obtidos da pesquisa são também ordenados pelo tipo de interrogação. Os documentos mais relevantes são os que foram obtidos pela execução da interrogação mais especifica e os documentos menos relevantes os que foram obtidos através da execução da interrogação mais geral.

Esta ordenação faz com que o sistema faça primeiro a pesquisa das respostas nos documentos recuperados pela execução das interrogações mais especificas e depois nos documentos recuperados pela execução das interrogações mais gerais.

Normalmente nos sistemas de perguntas e respostas, os documentos recuperados pelo sistema de recuperação de informação, são os documentos usados na extracção das respostas. Neste sistema isto não acontece pois as respostas são extraídas da interpretação semântico/pragmática os documentos de texto e não dos documentos de texto recuperados pelo motor de busca.

A informação que este módulo de recuperação de informação vai passar para os próximos módulos vão ser as DRSs dos documentos recuperados não os próprios documentos.

A pesquisa que o motor de busca faz é sobre os colecção de documentos de texto que já tinha sido indexada na fase de pré-processamento, no entanto, os documentos que vão ser usados nos módulos seguintes do sistema são as representações semânticas dos documentos.

Para isso foi feito um mapeamento entre os documentos de texto e as suas DRSs para que estes pudessem ser passados aos módulos seguintes. Assim, em vez do módulo de recuperação de informação passar os documentos de texto ao próximo módulo, são passadas as suas DRSs.

Para fazer o mapeamento entre os documentos relevantes foi criada uma pequena aplicação que recebe a identificação de um documento relevante e devolve a DRS do documento.

A arquitectura global do motor de busca pode ser vista na figura 3.5, onde se pode ver o uso da aplicação que faz o "mapeameamento"entre os documentos relevantes e as suas DRSs, bem como a sua interacção com o SINO.

Este motor de busca tem a característica de conseguir fazer a pesquisa nas várias secções do ficheiro de texto. Esta característica pode-se tornar bastante útil, pois pode ser feita uma pesquisa apenas numa das secções do ficheiro de texto em vez de pesquisar em todo o documento. Pesquisando apenas em algumas secções dos ficheiros de texto pode-se refinar mais a pesquisa, pois pode-se estar à procura de uma informação que esteja necessariamente numa determinada secção do ficheiro de texto, obtendo-se assim documentos mais relevantes para as perguntas.

Para que o motor de busca consiga fazer a pesquisa nas várias secções do documento de texto, este é necessário que esteja formatado com num formato especial do SINO. Esta formatação foi feita anteriormente no módulo de pré-processamento dos documentos de texto. Este pré-processamento dos documentos de texto é um processo bastante simples, inserindo apenas algumas etiquetas nos documentos.

Embora esta caracterista seja importante na recuperação dos documentos, nem sempre pode ser usada, visto que nem todos os documentos de texto podem ser

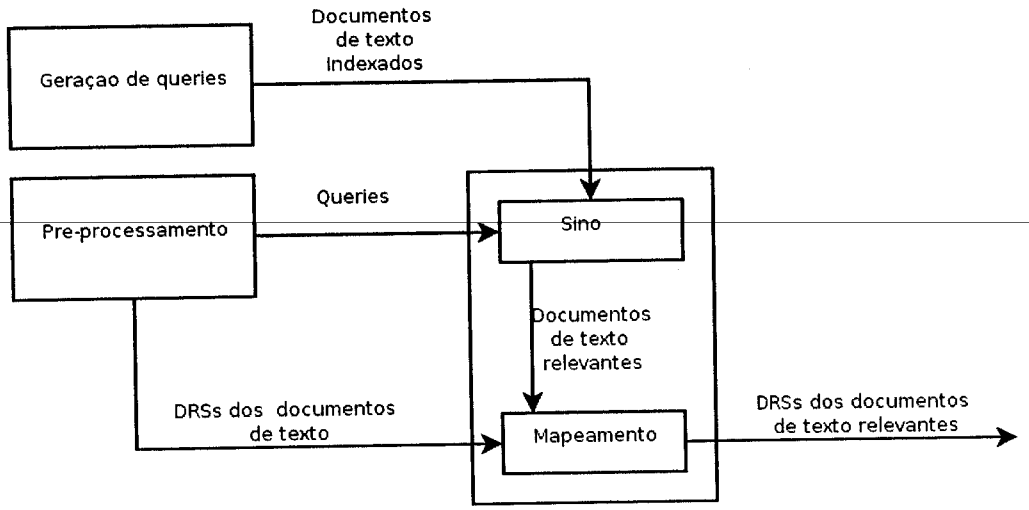

Figura 3.5: Arquitectura global do motor de busca.

separados por secções.

Apenas documentos de texto que já estejam de certa forma divididos por secções como por exemplo um artigo de um jornal que tem o titulo, o autor, o texto, ...é que podem é que podem ser tratados de forma a que o SINO possa fazer a pesquisa numa determinada secção do texto.

3.3.5 Interpretação semântico/pragmática dos documentos de texto

Depois de obtido o conjunto de DRSs dos documentos que possam ser relevantes para a pergunta é necessário fazer a interpretação semântico/pragmática destes documentos.

Para se obter um conjunto de factos de cada uma das frases dos documentos vai ser usada a ontologia para se obter o significado de cada uma das frases dos documentos de texto.

Este módulo recebe como *input* a representação semântica dos documentos relevantes que são compostos por várias DRSs(uma DRS para cada frase do documento).

O sistema analisa cada uma das DRSs das frases do documento que estão em *First Order Logic Form* e tenta interpretar cada uma das frases no contexto na

base de dados do documento(ontologia).

Para se conseguir fazer esta interpretação o sistema tenta encontrar a melhor explicação para a forma lógica de forma a que esta seja verdadeira na base de conhecimento usada na interpretação semântico/pragmática. Esta estratégia de interpretação é conhecida como *"interpretation as abduction"*[21]

A base de conhecimento para a interpretação semântico/pragmática é construída a partir da ontologia. A inferência usada nesta base de conhecimento usa "abduction" e restrições (GNU Prolog Finite Domain (FD) constraint solver).

As regras da base de conhecimento contêm informação para a interpretação de cada termo da frase na sua forma lógica como um termo de Prolog.

No seguinte exemplo pode-se ver a DRS gerada para a frase "O gato do João comeu o rato do Manuel."

```
drs(
    [def-A-m-s,def-B-m-s,
     def-C-m-s,def-D-m-s],
    [gato(A),rel(de,A,B),
     nome(B,'João'), comer(A,C),
     rato(C),rel(de,C,D),
     nome(C,'Manuel')]
  )
```

Nesta DRS pode-se ver a criação de 4 referentes e as suas relações. A interpretação semântico/pragmática da DRS desta frase vai dar origem á seguinte DRS:

```
drs(
    [def-A-m-s,def-B-m-s,
     def-C-m-s,def-D-m-s],
    [gato(A), owns(B,A), pessoa(B),
     nome(B,'João'), comer(A,C),
     rato(C),owns(D,C), pessoa(D),
     nome(C,'Manuel')]
  )
```

A interpretação de *rel(de, A, B)* como *owns(B,A)* é possível devido à existência da relação *owns* que relaciona pessoas com animais.

3.3.6 Extracção de respostas

Como o nome deste módulo indica, é aqui que vai ser feita a extracção das respostas exactas para cada pergunta. Para isso vão ser analisadas as interpretações semântico/pragmáticas das perguntas na forma de DRS e uma base de conhecimento representada numa forma lógica construída a partir dos 50 documentos de texto relevantes para a pergunta.

A forma como este sistema de as respostas são extraídas neste sistema é algo diferente de muitos outros sistemas de perguntas e respostas, sendo as respostas extraídas através dum processo de inferência.

O processo de inferência é feito através do uso do algoritmo de resolução do Prolog que tenta unificar os referentes da pergunta com factos extraídos dos documentos.

Esta unificação tem em conta o género e número de cada um dos referentes. O processo de inferência usa também algumas palavras que podem identificar o tipo da pergunta como "onde", "quando" e "quem" para identificar aquilo que está a ser perguntado.

Se por exemplo a pergunta for sobre um local de uma entidade especifica como na pergunta "Em que cidade se encontra a prisão de San Vittore?", o sistema vai tentar encontrar uma característica da entidade que é um local e que não é referida na pergunta.

Este sistema está dependente da qualidade da ontologia que foi inferida e da boa representação semântico/pragmática das frases e das perguntas.

Este sistema de perguntas e respostas, ao contrário de muitos outros não calcula qualquer valor de confiança para as respostas, sendo atribuído um valor máximo de 1 confiança a cada resposta. Esta característica do sistema é uma consequência da abordagem que foi feita para a extracção das respostas.

Esta valor de confiança significa que quando o sistema encontra uma resposta é porque tem a certeza que a resposta está correcta.

3.4 Aplicações do sistema

Este sistema pode ter inúmeras aplicações na vida real, podendo pesquisar respostas para perguntas em Língua Portuguesa em qualquer colecção de documentos de texto que estejam escritos em Língua Portuguesa. Uma vez que este sistema trabalha sobre um domínio aberto este passa a ser um bastante versátil.

3.4.1 Documentos da PGR

Quando este sistema começou a ser desenvolvido, estava disponível uma colecção de documentos de texto da Procuradoria Geral da Republica que já estavam a ser usados numa aplicação de recuperação de informação.

Para usar estes documentos no sistema de perguntas e respostas foi necessário fazer algum pré-processamento, no entanto não foi necessário passar por todos os passos do pré-processamento do sistema de perguntas e respostas, uma vez que tal foi feito anteriormente para a aplicação de recuperação de informação. O único pré-processamento que foi necessário efectuar foi a geração das DRSs do documentos de texto.

A indexação dos dos documentos de texto também não foi necessária visto que já tinha sido feita para a aplicação de recuperação de informação que também usava o SINO como motor de busca, sendo apenas necessário preparar o sistema de perguntas e respostas para que este saiba onde encontrar os documentos de texto para fazer as suas pesquisas.

Estes documentos de texto já estavam formatados de forma a poderem ser analisados pelo SINO, tendo mesmo cada documento de texto várias secções. Embora estas secções possam ser bastante úteis para a recuperação dos documentos relevantes, podem atrapalhar na geração das DRSs. Para isso foi necessário escolher e separar as secções dos documentos de texto que interessavam para a geração da sua representação semântica. Estes documentos tinham várias secções, como por exemplo: *descritores*, *data*, *texto*, *conclusões*, ... As secções que foram escolhidas para se gerarem as DRSs foram as secções *texto* e *conclusões*, pois são estas as secções dos documentos de texto que têm a informação que é importante para se poder responder à pergunta. Para se separar estas secções das outras foi usada a pequena ferramenta *Sed* junto com alguns *shell scripts* que ajudaram na criar os novos documentos de texto apenas com a informação necessária para gerar as DRSs.

Depois das DRSs dos documentos de texto estarem geradas e o sistema preparado para fazer a pesquisa nos documentos de texto da Procuradoria Geral da Republica, começaram-se a fazer algumas perguntas ao sistema sobre esta colecção de documentos. No entanto, devido à natureza dos documentos e à informação contida nos documentos, relevou-se bastante complicado encontrar informação nos documentos para responder a perguntas curtas e directas.

Com a informação contida nestes documentos de texto, torna-se difícil encontrar respostas para as perguntas devido aos temas dos textos e à forma como estes estão escritos.

No exemplo seguinte pode-se ver um excerto de um documento da PGR, podendo-se ver as várias *tags* do SINO a delimitar algumas secções do documento.

```
<!-- sino section DESCRITORES -->
CORPO DE BOMBEIROS
LEI ESPECIAL
SERVIÇO NACIONAL DE BOMBEIROS
FINANÇAS LOCAIS
COMPARTICIPAÇÃO
SUBSIDIO

<!-- end section --> <!-- sino section CONCLUSOES --> 1 - A Lei n
10/79, de 20 de Março, que ratifica o Decreto-Lei n 388/78, de 9 de
Dezembro, como lei especial que e, afasta a aplicação do artigo 16 da
Lei das Finanças Locais, aprovado pela Lei n 1/79, de 2 de Janeiro;

2 - A Lei n 10/79, de 20 de Março, permite a inclusão dos bombeiros
municipais e dos sapadores bombeiros, como beneficiarios dos subsidios
a atribuir, quer da colecta quer das dotações inscritas no orçamento
do Ministerio da Administração Interna, embora os destas ...
```

Neste exemplo pode-se ver também o uso de linguagem algo complexa devido ao tema dos assuntos que são tratados nos documentos.

Estes documentos estavam já a ser usados por uma aplicação *web* que fazia recuperação de informação, sendo esta aplicação basicamente um motor de busca sobre estes documentos. Tendo em conta que estes documentos já estavam a ser usados por uma aplicação de recuperação de informação e que os sistemas de perguntas e respostas podem também ser considerados como um sistema de recuperação de informação, este foi integrado na aplicação de recuperação de informação de forma a que se pudessem fazer perguntas em Língua Portuguesa sobre os documentos bem como fazer pesquisas normais de recuperação de informação sobre os mesmos documentos. Com o uso deste sistema de perguntas e respostas conseguiu-se aumentar as capacidades do sistema de pesquisa já existente.

3.4.2 Documentos do CLEF - Colecção do Publico

No decorrer deste projecto, surgiu uma oportunidade de participar na edição CLEF[8] de 2004 com este sistema de perguntas e respostas tarefa Monolingue PT-PT. Esta tarefa do CLEF, (QA@CLEF) tinha como objectivo responder a uma colecção de perguntas em Língua Portuguesa com base numa colecção de documentos de texto fornecidos pelo CLEF. Esta colecção de documentos de texto era

[8]Cross Language Evaluation Forum

composta por artigos do jornal "Publico" do anos de 1994 e 1995(51751 artigos de 1994 e 55070 artigos de 1995).

Estes documentos foram fornecidos pelo CLEF num conjunto de vários ficheiros. Estes ficheiro eram compostos por um conjunto de artigos do publico, estando cerca de 300 artigos do publico em cada ficheiro. O sistema de perguntas e respostas consegue trabalhar sobre qualquer tipo de ficheiro de texto, no entanto foi necessário separar os vários artigos dos ficheiros para que se tornasse mais simples a identificação dos artigos que serviram para responder às perguntas. A identificação dos documentos usados para responder às perguntas foi necessária devido às regras impostas pelo CLEF, sendo necessário indicar qual o documento usado para gerar a resposta. Tendo em conta esta necessidade, todos os artigos do publico foram separados, sendo criado um documento de texto para cada artigo.

Estes artigos do Publico são compostos por várias secções, tendo secções que identificam o documento, secções que indicam em que categoria o documento se insere, o autor, a data e o texto do artigo. Para que se pudesse usar todas as potencialidades do motor de busca utilizado neste sistema foi necessário formatar os documentos para que as secções estivessem delimitadas com as etiquetas do SINO para que este pudesse fazer a pesquisa nestas secções. Os documentos já tinham as secções separadas por etiquetas de XML, no entanto foi necessário acrescentar as etiquetas do SINO, sendo estas muito parecidas com as etiquetas de XML.

Depois dos documentos de texto estarem com o formato correcto, foram então indexados pelo SINO de forma a que pudesse ser feita a pesquisa nos documentos.

No exemplo seguinte pode-se ver o exemplo de um documento de texto do Publico já com as etiquetas do SINO:

```
<DOCNO>
<!-- sino section DOCNO -->
PUBLICO-19940701-057
</DOCNO>
<!-- end section -->
<DOCID>
<!-- sino section DOCID -->
PUBLICO-19940701-057
</DOCID>
<!-- end section -->
<DATE>
<!-- sino section DATE -->
19940701
</DATE>
<!-- end section -->
<CATEGORY>
```

```
<!-- sino section CATEGORY -->
Mundo
</CATEGORY>
<!-- end section -->
<AUTHOR>
<!-- sino section AUTHOR -->
JAF
</AUTHOR>
<!-- end section -->
<TEXT>
<!-- sino section TEXT -->
O protagonista
Filho de pescador
Quando a sua correligionária e presidente da câmara baixa do
parlamento, Takako Doi, anunciou o resultado da votação do nome para o
cargo de primeiro-ministro da segunda economia mais poderosa do mundo,
Tomiichi Murayama, 70 anos, não foi capaz de controlar o
coração. «Bateu desordenamente durante muitos minutos. Não estou
habituado a estas coisas», confessou.
Sobrancelhas muito espessas, esta será a característica física mais
forte de um homem seco, relativamente alto, afável, cujo ar
respeitável sai reforçado por cabelos abundantes e grisalhos.
Filho de um pescador da perfeitura de Oita, na ilha de Kyushu, no sul
do Japão, Murayama frequentou a Universidade Meiji de Tóquio, onde se
licenciou em economia e ciência política, e começou a trabalhar no
sindicato local dos pescadores, onde a ala esquerdista do PS continua
a ter grande força.
Casado, com duas filhas, foi vereador durante vários anos em Oita,
antes de se tornar deputado, em 1972. Durante os mais de 22 anos em
que foi sendo reeleito para a câmara baixa do parlamento foi sendo
reconhecido como um hábil negociador de consensos, nos bastidores,
dedicando-se aos sectores de pensões, bem estar social ou trabalho ...
```

Neste exemplo pode-se ver as várias secções do documento de texto, as *tags* de XML, bem como as *tags* são usadas pelo SINO para fazer a separação das secções. Neste exemplo pode-se também ver a simplicidade da linguagem utilizada no documento.

Depois da indexação dos documentos estar feita, foi necessário fazer a geração das DRSs para todos os documentos das colecção. Para fazer a esta geração foi necessário extrair o texto do artigo para que a DRS gerada fosse o melhor possível.

Para resolver este problema, foi então separada a secção que contém o texto actual dos documentos e gerados novos documentos de texto que contêm apenas o texto dos artigos. Sobre estes novos documentos de texto foram então geradas as DRSs, obtendo-se assim a representação semântica para a colecção de documentos do Publico.

Com a indexação dos documentos e a geração das DRSs feita, basta apenas preparar a aplicação para que esta esteja pronta a trabalhar com este novo conjunto de documentos. Para que a aplicação possa responder a perguntas em Língua Portuguesa sobre esta nova colecção de documentos, basta apenas indicar onde se encontram os documentos indexados e as DRSs dos documentos. Com esta informação é possível começar a fazer perguntas ao sistema e obter algumas respostas sobre esta colecção de documentos.

Com este pré-processamento dos documentos de texto fornecidos pelo CLEF, já é possível fazer perguntas em Língua Portuguesa ao sistema, no entanto para o CLEF isto não era suficiente. O CLEF fornecia um conjunto de 200 perguntas às quais o sistema devia responder sem qualquer intervenção humana. Estas perguntas eram fornecidas pelo CLEF num ficheiro com as várias perguntas e a identificação de cada uma das perguntas.

Para resolver este problema foi criada uma pequena aplicação que retira a pergunta em Língua Natural junto com a sua identificação, sendo depois esta pergunta passada ao sistema de perguntas e respostas. Depois do sistema de perguntas e respostas ter conseguido ou não obter uma resposta para a pergunta esta era passada novamente à aplicação que iria formatar a resposta no formato certo junto com a identificação da pergunta para que pudesse ser devolvida ao CLEF. Esta formatação da resposta era necessária visto que era um dos requisitos do CLEF. Todas as respostas tinham de estar identificadas e formatadas segundo um formato especificado pelo próprio CLEF.

No exemplo seguinte pode-se ver um extracto do documento com as perguntas fornecido pelo CLEF.

```
F PT PT 0001 Em que cidade se encontra a prisão de San Vittore?
F PT PT 0002 Onde era o campo de concentração de Auschwitz?
F PT PT 0003 Quem foi o autor de "Mein Kampf"?
F PT PT 0004 Qual é a capital da Rússia?
F PT PT 0005 Quem foi o primeiro presidente dos Estados Unidos?
F PT PT 0006 Como morreu Jimi Hendrix?
F PT PT 0007 Com quem se casou Michael Jackson?
F PT PT 0008 Em que género musical se distingue Michael Jackson?
D PT PT 0009 O que é a Mossad?
```

```
F PT PT 0010 Quantos crimes são atribuídos ao Monstro de Florença?
F PT PT 0011 Quantos desempregados há na Europa?
F PT PT 0012 Quantas religiões monoteístas há no mundo?
F PT PT 0013 Quantos judeus existem no mundo?
F PT PT 0014 Quantos detidos há no Corredor da Morte na Califórnia?
D PT PT 0015 O que é a UNICEF?
F PT PT 0016 Nomeie uma pessoa acusada de pedofilia.
D PT PT 0017 Quem é Jean-Bertrand Aristide?
```

Como se pode ver neste exemplo, cada pergunta tem uma identificação, bem como outras informações que para este sistema de perguntas e respostas não tem interesse. Esta a identificação deve ser usada nas respostas quando estas são enviadas novamente para o CLEF para identificar a que pergunta pertence a resposta.

Os documentos de texto fornecidos pelo CLEF são artigos do Publico, sendo textos não muito complexos nem muito técnicos uma vez que são lidos por todas as pessoas em jornais diários. Devido a esta característica dos textos, torna-se relativamente fácil encontrar respostas para perguntas simples, curtas e directas.

Embora este sistema estivesse desenhado para responder a este tipo de perguntas, conseguiu-se responder a perguntas um pouco mais complexas e maiores, isto devido à simplicidade de como os factos são apresentados nos documentos de texto. Os resultados obtidos por este sistema foram relativamente interessantes, visto que o sistema foi desenvolvido em pouco tempo e havia pouca experiência nesta área.

Este sistema conseguiu responder a cerca de 74 perguntas das 199 que foram feitas, estando 56 resposta certas e 18 incorrectas que corresponde a 28% de respostas correctas.

Com estes resultados podemos chegar à conclusão que a forma como os documentos de texto estão escritos e o seu conteúdo podem influenciar bastante os resultados obtidos pelo sistema. Verificou-se que se os documentos de texto forem simples e estiverem escritos numa linguagem simples o sistema consegue obter resultados bastante bons, no entanto se os documentos de texto forem escritos numa linguagem mais complexa e e mais técnica, o sistema tem algumas dificuldades em encontrar respostas, sendo bastante menos eficaz.

3.4.3 Aplicação concreta do sistema

Este sistema de perguntas e respostas foi integrado numa aplicação de recuperação de informação tradicional(ABC). Esta aplicação de recuperação de informação tem como objectivo obter documentos relevantes para uma certa inter-

rogação a partir de várias colecções de documentos e está implementada como um sistema WEB. Na figura 3.6 pode-se ver o resultado duma pesquisa no motor de busca ABC e o seu interface.

O objectivo da integração do sistema de perguntas e respostas nesta aplicação é adicionar algumas potencialidades ao sistema para que os utilizadores possam fazer perguntas em Língua Natural ao sistema sobre os vários conjuntos de documentos.

Outro objectivo desta integração é a criação de um interface para o sistema de perguntas e respostas implementado de forma a que este possa ser facilmente testado em várias situações.

O sistema foi integrado de duas formas diferentes, permitindo fazer as pesquisas das respostas nas várias colecções de documentos ou num texto introduzido pelo utilizador. Esta ultima forma de fazer a pesquisa serve principalmente para facilitar os testes do sistema com novos textos sem ter que estar a adicionar os textos numa das colecções de documentos. Na figuras 3.7 e 3.8 podem-se ver respectivamente os interfaces para o modo que faz a pesquisa da resposta nas colecções de documentos e para o modo que faz a pesquisa no texto inserido pelo utilizador.

Como se pode ver nas figuras 3.7 e 3.8 torna-se bastante simples fazer perguntas ao sistema, tanto quando se pretende pesquisar a resposta na colecção de documentos ou no texto inserido pelo utilizador. Com este interface, qualquer utilizador que esteja familiarizado com a WEB pode fazer perguntas ao sistema, ficando assim o sistema disponível a qualquer utilizador.

Os resultados apresentados pelo sistema são também bastante fáceis de interpretar para os dois modos dos sistemas, sendo apresentados de uma forma clara e simples.

São também apresentadas algumas informações para além da resposta obtida pelo sistema, como é o caso do "log" do sistema de inferência e a lista de documentos relevantes que foram obtidos pelo pelo sistema de recuperação de informação.

Estas informações destinam-se principalmente para analisar os resultados obtidos. Com esta informação torna-se mais fácil encontrar as causas para alguns problemas do sistema, uma vez que o utilizador pode consultar o "log" do sistema de inferência de respostas, ver toda a lista de ficheiros relevantes que foram recuperados pelo sistema de recuperação de informação e consultar o conteúdo de cada um dos ficheiros recuperados. Nas figuras 3.9 e 3.10 pode-se respectivamente o resultado da procura duma resposta nas colecções de documentos e num texto introduzido pelo utilizador.

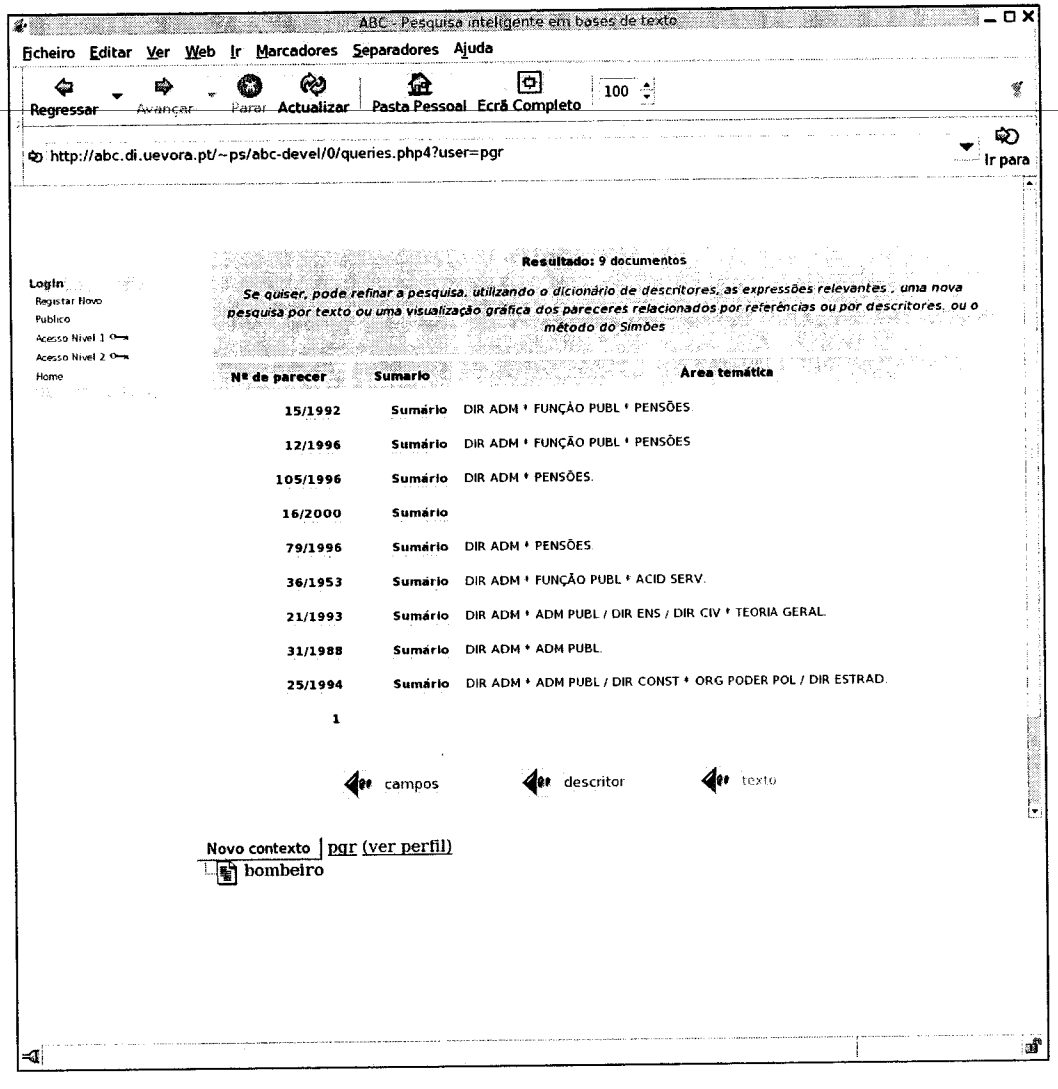

Figura 3.6: Resultado duma pesquisa no motor de busca ABC.

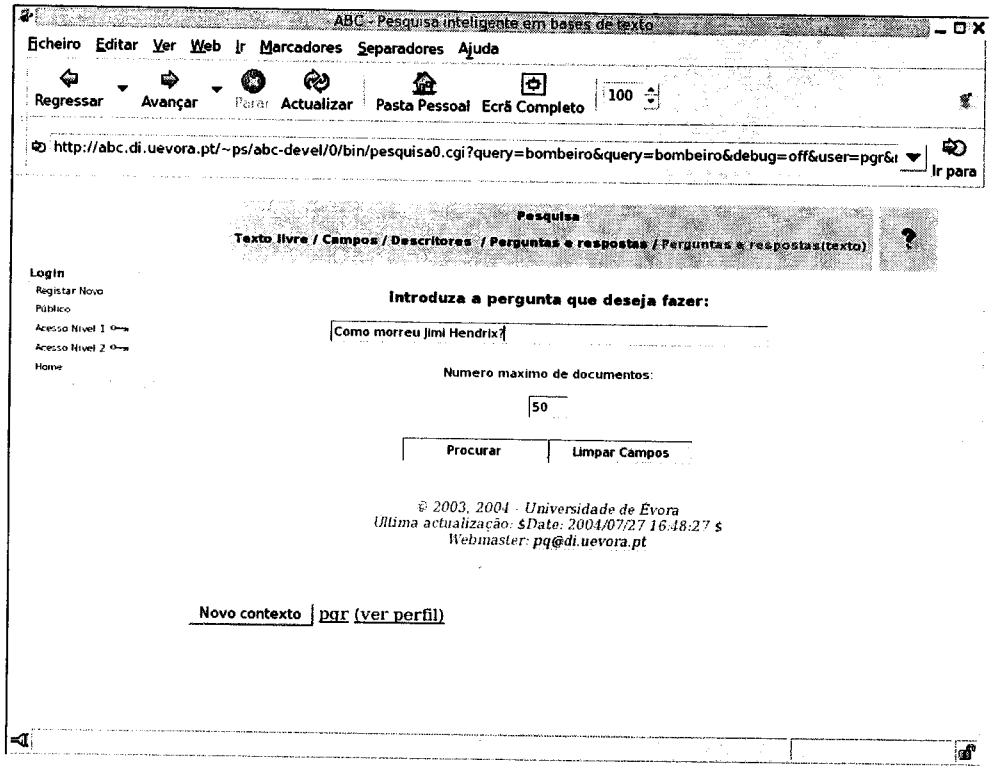

Figura 3.7: Interface do sistema que pesquisa nas colecções de documentos.

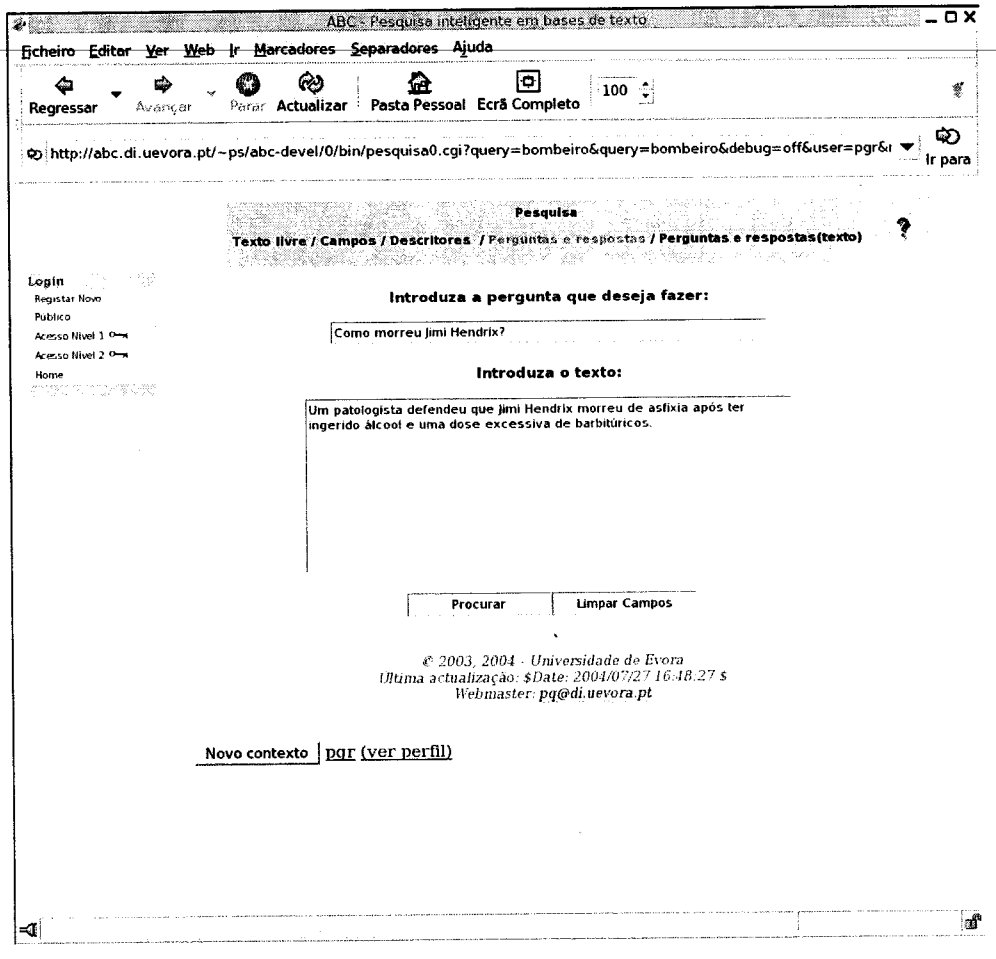

Figura 3.8: Interface do sistema que pesquisa no texto inserido pelo utilizador.

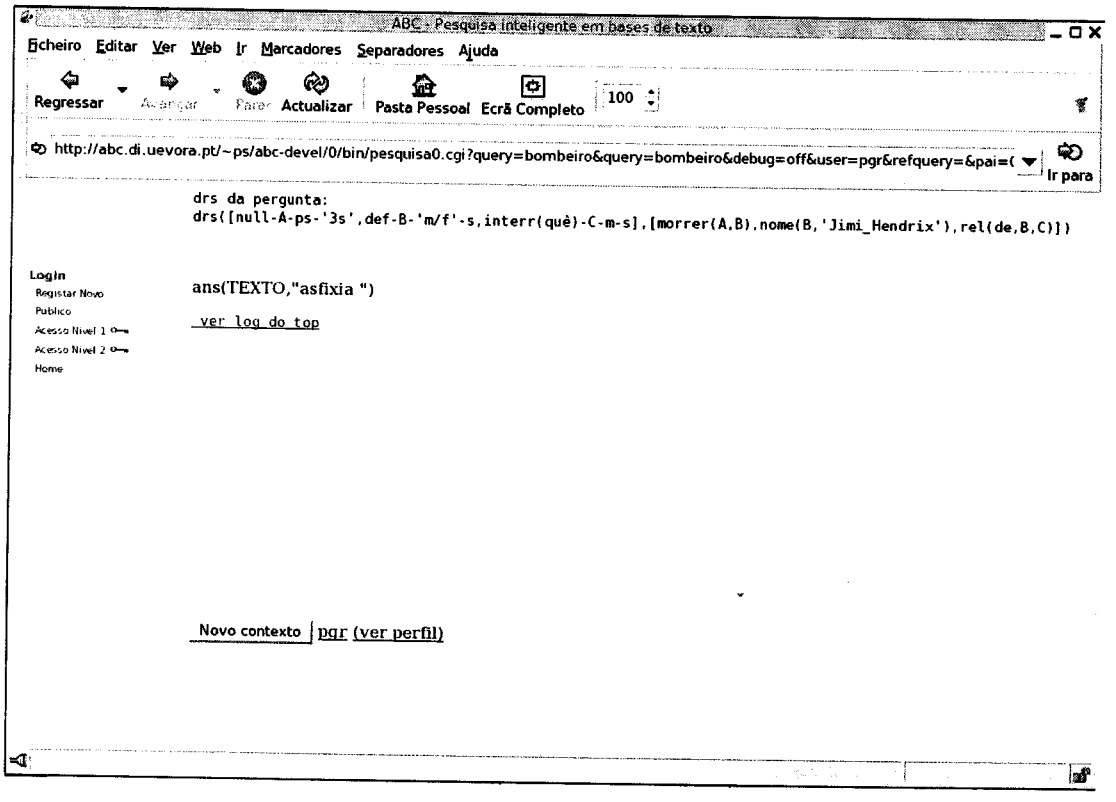

Figura 3.9: Resultado do sistema que pesquisa nas colecções de documentos.

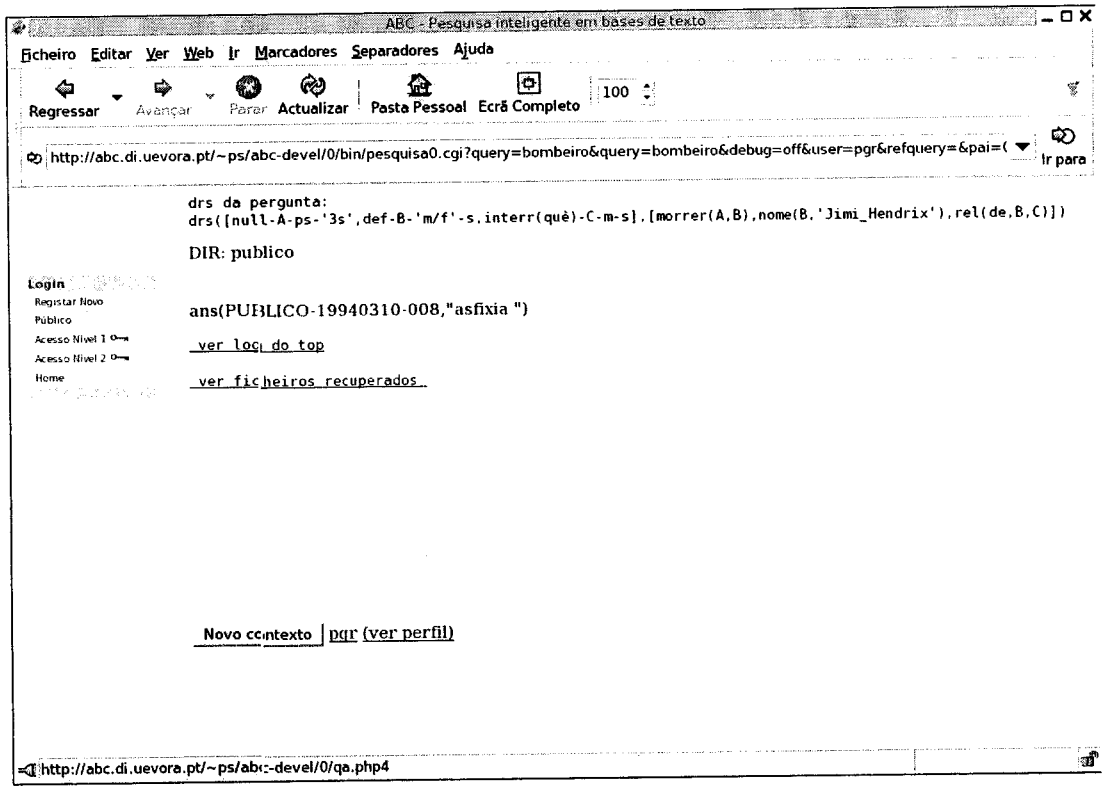

Figura 3.10: Resultado do sistema que pesquisa no texto inserido pelo utilizador.

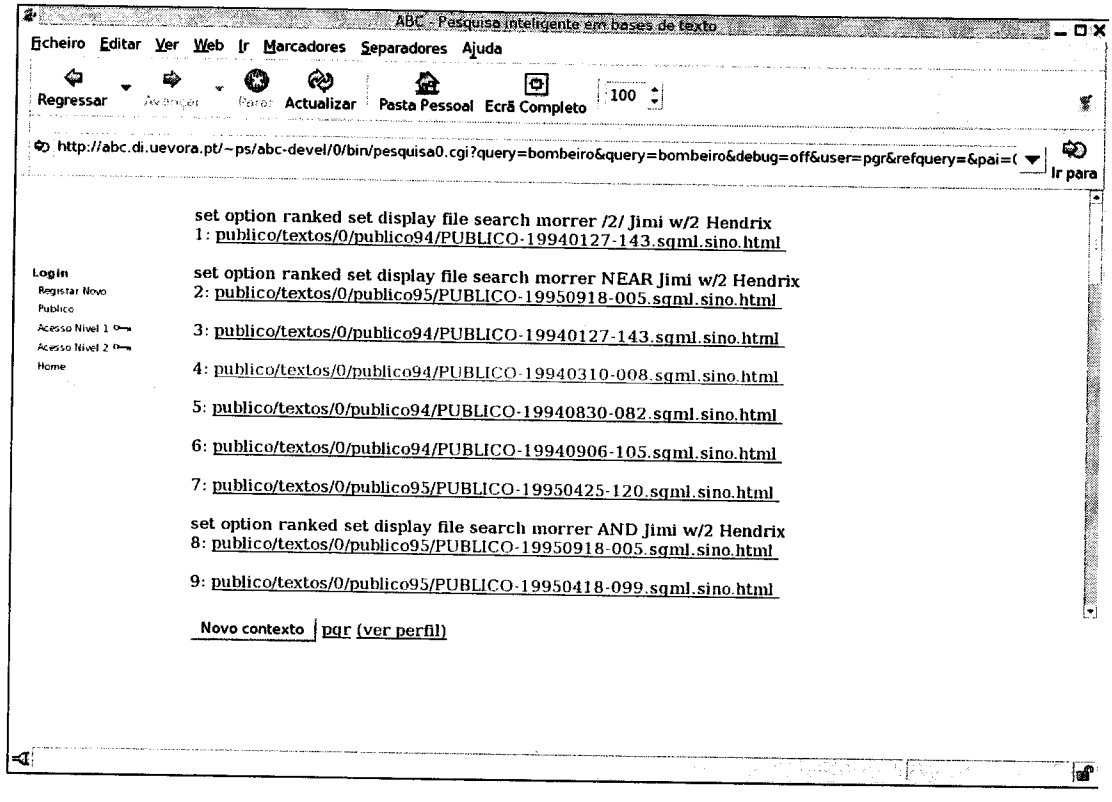

Figura 3.11: Resultado do sistema que pesquisa no texto inserido pelo utilizador.

Como se pode ver na figura 3.9, o utilizador pode facilmente ver o "log" do sistema de inferência das respostas(top) ou consultar a lista de documentos relevantes que foram recuperados. Nesta imagem pode-se também ver que a resposta é apresentada de uma forma clara e simples, sendo também apresentado qual o documento que foi usado para construir a resposta.

Na figura 3.11 pode-se ver a lista de documentos relevantes que foram recuperados bem como a interrogação que foi usada para recuperar os vários conjuntos de documentos. O utilizador ao consultar esta lista pode também consultar o conteúdo de qualquer documento recuperado através dum simples clique no respectivo "link".

Esta integração no sistema de perguntas e respostas no ABC tem-se revelado bastante útil, principalmente na identificação de alguns problemas ainda existentes.

3.5 Comparação com o estado da arte

Este sistema de perguntas e respostas tem muitas semelhanças com outros sistemas já implementados para outras línguas como o Inglês. Comparativamente a esses sistemas de perguntas e respostas este sistema tem uma arquitectura global bastante semelhante, pois, tal como a maior parte dos sistemas de perguntas e respostas estão divididos por vários módulos, estando cada fase do processo de responder às perguntas atribuída a cada um dos módulos.

Normalmente estes sistemas de perguntas e respostas têm um módulo de processamento das perguntas, um módulo que faz uma geração de interrogação, um módulo que faz a recuperação dos documentos relevantes para as perguntas e por fim um outro módulo que faz a extracção das respostas.

Tal como estes sistemas de perguntas e respostas, este sistema contém os mesmos módulos, no entanto a forma como foram implementados e principalmente as teorias que suportam a metodologia usada na implementação de alguns módulos difere muito dos outros sistemas, sendo principalmente na parte do processamento das perguntas, dos documentos e da extracção das respostas que se verifica mais esta diferença.

A principal característica deste sistema é o uso da semântica e a forma como é representada no processamento das perguntas, dos documentos e na extracção da resposta. Esta representação semântica assenta sobre a DRT.

Embora o uso desta representação semântica seja a principal diferença para os outros sistemas de perguntas e respostas, existe pelo menos um outro sistema de perguntas e respostas que usa também a DRT. Este sistema de perguntas e respostas chama-se QED e está descrito na secção 2.3.3.

Este sistema é bastante parecido com o nosso sistema de perguntas e respostas, no entanto tem algumas diferenças principalmente na forma como as perguntas são extraídas. No QED a extracção das respostas assenta também sobre o *match* das DRSs das perguntas com as DRSs dos documentos, no entanto usa algumas consultas ao WordNET para ter mais *certeza* na resposta.

Outra diferença do QED é que devolve várias respostas para cada pergunta sendo depois estas perguntas ordenadas com algumas pesquisas no Google. Ao contrario do QED, no nosso sistema de perguntas e respostas devolve apenas uma resposta para cada pergunta.

Os resultados do sistema desenvolvido comparados com os resultados obtidos QED foram bastantes bons, conseguindo-se resultados bastante iguais. Tendo em conta estes resultados e que este sistema de perguntas e respostas pode ser considerado ainda como um protótipo os resultados são muito bons e bastante motivadores.

3.6 Conclusão

Este sistema de perguntas e respostas representa uma primeira abordagem para um sistema de perguntas para a Língua Portuguesa.

Este sistema usa técnicas de processamento de língua natural para criar uma base de conhecimento a partir da colecção de documentos. As perguntas são analisadas por ferramentas de processamento de língua natural e são feitas inferências sobre a base de conhecimento para se tentar obter uma resposta correcta para a pergunta.

A ideia inicial de criar uma única base de conhecimento de grandes dimensões a partir dos factos extraídos de todos os documentos não se tornou viável devido a problemas de complexidade computacional. Estes problemas levaram à criação de processos de recuperação de informação sobre as perguntas para diminuir a complexidade da base de conhecimento. No entanto, o sistema de recuperação de informação mostrou alguns problemas que levaram à incapacidade do sistema de perguntas e respostas de obter respostas para muitas perguntas.

A ontologia usada foi também um grande problema do sistema dando origem a muitas respostas erradas. O problema relacionado com a ontologia é a forma como deve ser criada a ontologia. Este sistema foi desenvolvido para trabalhar com documentos de domínio geral. O uso de documentos de domínio geral pode-se tornar um grande problema na geração de uma boa ontologia geral, dado não se saber bem como criar uma ontologia para um domínio geral.

Como trabalho futuro pretende-se desenvolver novas ideias para resolver parcialmente estes problemas. Com o uso de novas estratégias de implementação pode ser possível ter uma única base de conhecimento e o uso de ontologias existentes junto com o WordNet pode melhorar a qualidade da ontologia final.

Pretende-se também explorar o problema da resolução de anáforas em várias fases do sistema.

Capítulo 4

Avaliação do Sistema

Neste capítulo é feita uma descrição e uma avaliação crítica dos resultados obtidos pelo sistema de perguntas e respostas desenvolvido no Departamento de Informática da Universidade de Évora.

Nas próximas secções são mostrados os resultados obtidos, comparações com outros sistemas, exemplos de perguntas que o sistema teve sucesso a responder, alguns problemas que houve no sistema bem como perguntas que tiveram esses mesmos problemas e algumas formas de resolver esses problemas.

4.1 Resultados

Este sistema foi usado num conjunto de 199 perguntas em Língua Portuguesa que faziam parte da colecção de perguntas disponibilizadas pelo CLEF na tarefa QA@CLEF-2004.

Neste conjunto de perguntas o sistema obteve 56 respostas correctas, 18 respostas inexactas e 125 respostas erradas, obtendo uma percentagem de respostas certas de 28.1% e um valor de confiança de 0.243. Se o valor a percentagem de respostas correctas for calculado sobre as respostas inexactas este valor sobe para 37.2%.

Estes resultados são bastante interessantes, mostrando o grande potencial desta abordagem aos sistemas de perguntas e respostas. No estando, estes resultados mostra também um dos grandes problemas do sistema, identificando 125 perguntas como não tendo resposta nos documentos enquanto que apenas 9 perguntas não tinham resposta.

No *Apêndice B* estão todas as perguntas feitas pelo QA@CLEF-2004 e todas as respostas dadas pelo nosso sistema.

Run Name	Task	Respostas	Correctas	Erradas	Inexactas	Não suportadas
PTUE041ptpt	PT→PT	199	56	125	18	0
sfnx041ptpt	PT→PT	199	22	165	8	4
sfnx041ptpt	PT→PT	199	30	154	10	5

Tabela 4.1: Resultados dos sistemas que usaram a língua Portuguesa como língua alvo

Run Name	Task	Respostas	Correctas	Erradas	Inexactas	Não suportadas
uams041ennl	EN→NL	200	70	122	7	1
uams041nlnl	NL→NL	200	88	98	10	4
uams042nlnl	NL→NL	200	91	97	10	2

Tabela 4.2: Resultados dos sistemas que usaram a língua Holandesa como língua alvo

4.2 Resultados de outros sistemas

Nesta secção são mostrados os resultados obtidos pelos vários sistemas de perguntas e respostas participantes no QA@CLEF-2004, incluindo o outro sistema de perguntas e respostas que participou na tarefa monolingue PT-PT do QA@CLEF.

Nas seguintes tabelas são apresentados o número de respostas apresentados em cada "run" de cada sistema, o número de respostas correctas, o número de respostas incorrectas, o número de respostas inexactas e o número de resposta não suportadas.

Como se pode pelos resultados na tabela 4.2, o nosso sistema(PTUE041ptp) foi o que obteve melhores resultados a usar a língua Portuguesa como língua alvo. Ao analisar os resultados dos sistemas para as outras línguas chega-se também à conclusão que o nosso sistema teve um desempenho bastante positivo em relação aos outros sistemas, sendo o sétimo sistema a obter mais respostas correctas num total de 48 "runs".

Run Name	Task	Respostas	Correctas	Erradas	Inexactas	Não suportadas
ILCP-QA-ITIT	IT→IT	200	51	117	29	3
irst041itit	IT→IT	200	56	131	11	2
irst021itit	IT→IT	200	44	147	9	0

Tabela 4.3: Resultados dos sistemas que usaram a língua Italiana língua alvo

Run Name	Task	Respostas	Correctas	Erradas	Inexactas	Não suportadas
gine041bgfr	BG→FR	200	13	182	5	0
gine041defr	DE→FR	200	27	162	11	0
gine041enfr	EN→FR	200	16	171	13	9
gine041esfr	ES→FR	200	25	166	9	0
gine041frfr	FR→FR	200	26	160	14	0
gine041itfr	IT→FR	200	23	166	11	0
gine041nlfr	NL→FR	200	17	171	12	0
gine041ptfr	PT→FR	200	22	170	8	0
gine042bgfr	BG→FR	200	13	180	7	0
gine042defr	DE→FR	200	32	155	13	0
gine042enfr	EN→FR	200	25	165	10	0
gine042esfr	ES→FR	200	30	164	6	0
gine042frfr	FR→FR	200	42	147	11	0
gine042itfr	IT→FR	200	27	165	8	0
gine042nlfr	NL→FR	200	26	158	16	0
gine042otfr	PT→FR	200	25	166	9	0

Tabela 4.4: Resultados dos sistemas que usaram a língua Francesa como língua alvo

Run Name	Task	Respostas	Correctas	Erradas	Inexactas	Não suportadas
aliv041eses	ES→ES	200	63	130	5	2
aliv042eses	ES→ES	200	65	129	4	2
cole041eses	ES→ES	200	22	178	0	0
inao041eses	ES→ES	200	45	145	5	5
inao042eses	ES→ES	200	27	152	6	5
mira041eses	ES→ES	200	18	147	7	1
talp041eses	ES→ES	200	48	150	1	1
talp042eses	ES→ES	200	52	143	3	2

Tabela 4.5: Resultados dos sistemas que usaram a língua Espanhola como língua alvo

Run Name	Task	Respostas	Correctas	Erradas	Inexactas	Não suportadas
bgas041bgen	BG→EN	200	26	168	5	1
dfki041deen	DE→EN	200	47	151	0	2
dltg041fren	FR→EN	200	38	155	7	0
dltg042frrn	FR→EN	200	29	164	7	0
edin041deen	DE→EN	200	28	166	5	1
edin041fren	FR→EN	200	33	161	6	0
edin042deen	DE→EN	200	34	159	7	0
edin042fren	FR→EN	200	40	153	7	0
hels041fren	FR→EN	193	21	171	1	0
irst041iten	IT→EN	200	45	146	6	3
irst042iten	IT→EN	200	35	158	5	2
lire041fren	FR→EN	200	22	172	6	0
lire042fren	FR→EN	200	39	155	6	0

Tabela 4.6: Resultados dos sistemas que usaram a língua Inglesa como língua alvo

Run Name	Task	Respostas	Correctas	Erradas	Inexactas	Não suportadas
dfki041dede	DE→DE	197	50	143	1	3
FUHA041dede	DE→DE	197	67	128	2	0

Tabela 4.7: Resultados dos sistemas que usaram a língua Alemã como língua alvo

4.3 Exemplos

Nesta secção são mostradas algumas perguntas da colecção de perguntas do CLEF para as quais o sistema conseguiu encontrar as respostas junto com as representações semânticas das perguntas e das frases que continham a resposta.

- ''Como morreu Jimi Hendrix?''

DRS da pergunta:

```
drs(
    [def-A-'m/f'-s,interr(quê)-B-m-s],
    [morrer(A),
     nome(A,'Jimi_Hendrix'),
     rel(de,A,B)]
  )
```

Texto com a resposta:

''Um patologista defendeu que Jimi Hendrix morreu de asfixia após ter ingerido álcool e uma dose excessiva de barbitúricos''.

DRS do texto:

```
drs([indef-A-m-s, def-B-'m/f'-s,
     def-C-f-s, def-D-m-s, indef-E-f-s],
    [patologista(A), defender(A,B),
     nome(B,'Jimi_Hendrix'), morrer(B),
     rel(de,B,C), asfixia(C),
     rel(após,C,D), ingerir(D),
     álcool(D) ,dose(D), excessivo(D),
     rel(de,D,E), barbitúrico(E)])
```

Resposta dada pelo sistema:

```
''asfixia''
```

- ''Quantos desempregados há na Europa?''

DRS da pergunta:

```
drs(
    [interr(quanto)-A-m-p, def-B-f-s],
    [desempregar(A,C), haver(C),
     rel(em,C,B),
     nome(B,'Europa')]
    )
```

Texto com a resposta:

```
''23 milhões de desempregados na Europa - A taxa de desemprego
nos países industrializados deverá crescer 0,1 por cento até
ao fim do ano, atingindo 8,6 por cento da população activa,
ou seja, 33 milhões de pessoas, 23 milhões dos quais na
Europa ocidental, sublinha um relatório da Organização
Internacional do Trabalho (OIT) ontem divulgado em Genebra.''
```

DRS do texto:

```
drs([null-A-null-null, def-B-f-s, def-C-f-s,
     def-B-f-s, null-D-null-null, def-E-m-s,
     def-F-m-s, null-G-null-null, def-H-f-s, def-K-f-s,
     null-L-M-N, def-O-f-s, indef-P-m-s, def-Q-f-s],
    [num(A,'23'), milhão(A), rel(de,A,T), desempregar(T,U),
     rel(em,T,B), nome(B,'Europa'), taxa(B), rel(de,B,C),
     desemprego(C), país(B), industrializar(B,V), crescer(V,D),
     num(D,'0,1'), por_cento(D),rel(a,D,E),fim(E),rel(de,E,F),
```

```
      ano(F), atingir(E,G), num(G,'8,6'), por_cento(G),
      rel(de,G,H), população(H), activo(H), ou_seja(H),
      num(H,'33'), milhão(H), rel(de,H,K), pessoa(K),
      num(K,'23'), milhão(K), rel(de,K,L),
      o_qual(L), rel(em,L,O), nome(O,'Europa'), ocidental(O),
      sublinhar(O,P), relatório(P), rel(de,P,Q),
      nome(Q,'Organização_Internacional_do_Trabalho'),
      nome(Q,'OIT'),ontem(Q),divulgar(Q,Q),
      rel(em,Q,Q),nome(Q,'Genebra')]
    )
```

Resposta dada pelo Sistema:

``23 milhão''

- ``Com quem se casou Michael Jackson?''

DRS da pergunta:

```
drs(
    [interr(quem)-A-'m/f'-'s/p',def-B-m-s],
    [rel(com,C,A),se(A),
     casar(A,B),nome(B,'Michael_Jackson')]
  )
```

Texto com a resposta:

``Lisa Marie Presley confirma casamento com Michael
Jackson
A senhora Presley-Jackson
A filha de Elvis Presley, Lisa Marie, 26
anos, confirmou o seu casamento há 11 semanas atrás
com o cantor Michael Jackson, 35 anos, num comunicado
enviado pela MJJ Productions, encarregada na
segunda-feira das relações públicas do cantor para
os principais «media».''

DRS do texto:

```
drs(
    [def-A-f-s, def-B-m-s, def-C-m-s, def-D-'m/f'-s,
     null-E-m-s, def-F-m-s, null-G-'F'-'P',
     def-J-m-s, indef-K-m-s,def-D-f-s],
    [nome(A,'Lisa_Marie_Presley'), confirmar(A,B),
     casamento(B), rel(com,B,C), nome(C,'Michael_Jackson_A'),
     senhora(C), nome(C,'Presley-JacksonPresley-Jackson_AA'),
     filho(C), rel(de,C,D), nome(D,'Elvis_Presley'),
     nome(D,'Lisa_Marie'), ano(D,'26'), confirmar(D,E),
     seu(E,F), casamento(F), rel(há,F,G), semana(G,'11'),
     atrás(G), rel(com,G,J), cantor(J), nome(J,'Michael_Jackson'),
     ano(J,'35'), rel(em,J,K), comunicado(K), enviar(D,K),
     nome(D,'MJJ'),np(D)]
    )
```

Resposta dada pelo sistema:

```
''Lisa Marie Presley''
```

- ''Onde é o hospital Júlio de Matos?''

DRS da pergunta:

```
drs(
    [interr(que)-A-B-C,def-D-m-s],
    [local(A),ser(A,D),
     hospital(D),nome(D,'Júlio_de_Matos')]
    )
```

Texto com a resposta:

Tem sede provisória no Hospital Júlio de Matos, em Lisboa, mais precisamente nos Serviços de Psicoterapia Comportamental (Av. do Brasil, 53 -- 1700 Lisboa).

DRS do texto:

```
drs(
     [null-G-pr-'3s', def-H-f-s,
      def-I-m-s, def-J-m-s],
     [ter(G,H), sede(H), provisório(H), rel(em,H,I),
      nome(I,'Hospital_Júlio_de_Matos'), rel(em,I,J),
      nome(J,'Lisboa'), mais(J), precisamente(J),
      nome(J,'Serviços_de_Psicoterapia_Comportamental'),
      nome(J,'Av._do_Brasil'), num(J,'53'),
      num(J,'1700'),nome(J,'Lisboa')]
)
```

Resposta dada pelo sistema:

```
Lisboa
```

4.3.1 Processo de inferência

Nesta secção é possível ver em detalhe o processo de inferência da resposta para uma pergunta, tomando como exemplo a pergunta "Como morreu Jimi Hendrix", a sua DRS e a DRS do texto com a resposta:

- **DRS da pergunta:**

```
drs(
     [def-A-'m/f'-s,interr(quê)-B-m-s],
     [morrer(A),
      nome(A,'Jimi_Hendrix'),
      rel(de,A,B)]
   )
```

- **DRS do documento:**

```
drs([indef-A-m-s, def-B-'m/f'-s,
     def-C-f-s, def-D-m-s, indef-E-f-s],
    [patologista(A), defender(A,B),
     nome(B,'Jimi_Hendrix'), morrer(B),
     rel(de,B,C), asfixia(C),
     rel(após,C,D), ingerir(D),
     álcool(D) ,dose(D), excessivo(D),
     rel(de,D,E), barbitúrico(E)])
```

Através destes exemplos consegue-se ver facilmente que o processo de inferência vai unificar a DRS da pergunta com a DRS do texto. Neste caso as condições da DRS da pergunta e as condições DRS do texto vão unificar e obter uma resposta.

- **Condições da DRS da pergunta**

```
morrer(A), nome(A,'Jimi_Hendrix'), rel(de,A,B)]
```

- **Condições da DRS do documento**

```
nome(B,'Jimi_Hendrix'), morrer(B), rel(de,B,C), asfixia(C)
```

Através dos referentes da DRS da pergunta, sabe-se que está a ser perguntado algo sobre o referente "B"(interr(quê)-B-m-s) que está representado nas condições da DRS por "rel(de,A,B)". Nas condições da DRS da frase existe a mesma condição ("rel(de,B,C)") que está associada ás condições que já foram unificadas e a uma nova condição que será a resposta para a pergunta: "asfixia(C)".

4.4 Problemas

Foi feito um estudo sobre as perguntas da colecção CLEF que não obtiveram resposta para se tentar encontrar as causas que levaram o sistema a não encontrar respostas para as perguntas. Desse estudo foi concluído que o sistema não obteve respostas para as perguntas devido ao sistema de recuperação de informação, à ontologia e ao processo de inferência.

Com este estudo ficou demonstrado que o problema maior do sistema se encontra no sistema de recuperação de informação, pois em 125 perguntas para as quais não foram encontradas respostas, não foram encontrados documentos relevantes para 80 perguntas.

4.4.1 Problemas com o SINO

O sistema de recuperação de informação usado para obter um conjunto de documentos relevantes para a pergunta de forma a reduzir a complexidade da base de conhecimento não conseguiu na maior parte das perguntas encontrar os documentos mais relevantes para a pergunta.

O sistema de recuperação de informação não conseguiu encontrar os documentos relevantes para bastantes perguntas principalmente devido às seguintes causas:

- Problemas de sinónimos
- Pesquisa de palavras não relevantes

As causas mais comuns para o SINO não encontrar os documentos relevantes é o uso de interrogações pouco relevantes para a pergunta que usam palavras da pergunta quando devia ser usado alguns sinónimos e o uso de interrogações com palavras da pergunta deviam ser ignoradas.

De seguida são mostrados alguns exemplos de perguntas que tiveram alguns problemas na recuperação de documentos relevantes para a pergunta:

- Na pergunta "Quem foi o primeiro presidente dos Estados Unidos?", o SINO estava à procura de "presidente dos Estados Unidos", no entanto, a resposta para a pergunta encontra-se no seguinte bloco de texto:

  ```
  ''John Kennedy Junior vai lançar um magazine político,
  «George», em homenagem ao primeiro presidente norte-americano,
  George Washington".''
  ```

  Neste exemplo devia ser usada uma interrogação que substitui-se "presidente dos Estados Unidos" pelo sinónimo "presidente norte-americano", conseguindo-se assim encontrar o documento relevante.

- Na pergunta "Onde fica o Museu do Hermitage?", não foram encontrados documentos relevantes pois na interrogação que é feita ao SINO é usado o verbo "ficar", ou seja, vai ser procurado um documento que tenha a palavra "ficar". Este é um verbo que ou devia ser ignorado na interrogação ou substituído por um sinónimo, pois a resposta foi encontrada no seguinte bloco de texto:

  ```
  ''Não existe nenhuma tábua do pintor em Portugal -- o seu
  espólio está espalhado em museus como o Hermitage de São
  Petersburgo, o San Mateo de Pisa, a Pinacoteca Nacional de
  ```

```
Cagliari, o Naradowe de Varsóvia, o Mable and Ringling de
Sarasota (Flórida), o Szépmuvészeti de Budapeste, o Herzog
Anton Ullrich de Braunschweig ou o Preussisher Kultburbesitz
de Berlim.''
```

Como se pode ver, neste bloco de texto o verbo "ficar" não existe, devendo este ser ignorado de forma a que o sistema de recuperação de informação consiga encontrar o documento.

- Na pergunta "Qual o acrónimo da Amnistia Internacional?" foi construída uma interrogação com a palavra "acrónimo" no entanto esta palavra devia ser ignorada pois é uma palavra que provavelmente não vai aparecer no documento de texto junto com a resposta. Como se pode ver no bloco de texto que contêm a resposta correcta para a pergunta, a palavra "acrónimo" não aparece:

```
''O grupo de trabalho «Angola» da Amnistia Internacional (AI),
já contestou a «alegada falta de poderes» do senador
democrata-cristão.''
```

Estes três exemplos demonstram bem as principais causas que levaram o SINO a não obter documentos relevantes para as perguntas.

Estas causas foram obtidas através dum estudo feito sobre as primeiras 20 perguntas da colecção de perguntas do CLEF que não obtiveram resposta.

Neste estudo chegou-se à conclusão que não foram obtidos documentos relevantes para 8 perguntas devido a problemas de sinónimos e para 9 perguntas por pesquisa de palavras não relevantes para a pergunta.

4.4.2 Problemas com a Ontologia

Um dos outros problemas para o sistema não encontrar uma resposta para muitas perguntas é a qualidade da ontologia, pois o processo de inferência da resposta depende muito da ontologia. Neste tipo de sistemas é muito importante saber o que são locais, pessoas, datas, sinónimos, ...

Um dos problemas graves da ontologia é a falta de relações sinonimiais, permitindo assim saber quais as entidades que são sinónimos. Com estas relações no ontologia o sistema torna-se mais eficaz ao fazer a inferência das respostas.

No exemplos seguintes pode-se ver algumas perguntas que o sistema teve problemas a encontrar uma resposta devido a problemas relacionados com a ontologia:

- Na pergunta "Em que cidade se encontra a prisão de San Vittore?", se a ontologia não tivesse a informação da relação entre "cidade" com uma classe "locais", então o sistema nunca conseguiria obter uma resposta para a pergunta.

- A resposta para a pergunta "Qual a localização de Tipaza?" encontrava-se no seguinte bloco de texto:

  ```
  O último terramoto que fez tremer a Argélia, em 1989, na região
  de Tipaza, a 70 quilómetros de Argel, a capital, matou 30
  pessoas e deixou centenas desalojadas.
  ```

  Para se conseguir fazer uma boa inferência sobre este bloco de texto, é preciso que a ontologia tenha a informação de que "a 70 quilómetros de Argel".

  Neste caso, a ontologia não tinha essa informação levando a que não se conseguisse responder à pergunta.

- Na pergunta "Em que cidade americana se encontra o Museu Warhol?" a ontologia deve ter a informação que "se encontra" pode estar de alguma forma relacionada com "inaugura-se", isto para que se consiga fazer a inferência da resposta a partir do seguinte bloco de texto:

  ```
  Na pequena cidade americana de Pittsburgh inaugura-se hoje o
  museu Andy Warhol, o maior do mundo dedicado a um só artista.
  ```

4.4.3 Problemas no processo de inferência

O processo de inferência das respostas também limitou um pouco o sistema, sendo também responsável por algumas perguntas sem respostas.

Uma causa que levou o processo de inferência a falhar na geração de algumas resposta foi a geração errada de algumas DRS's das perguntas como dos documentos de texto. As DRS's dos documentos de texto foram aquelas onde se verificaram mais erros, isto devido à complexidade de algumas frases.

Uma outra causa que levou o processo de inferência a falhar foi a grande quantidade de anáforas contidas nos documentos de texto.

4.5 Resolução de problemas

Estes problemas podem ser resolvidos várias formas. Em relação ao sistema de recuperação de informação, as interrogações ao SINO devem ser geradas tendo em conta um dicionário de sinónimos, sendo usados os diversos sinónimos para cada palavra da interrogação, isto para que a interrogação não seja tão especifica. Deve também ser usada uma lista de palavras que não devem ser usadas nas interrogações. Estas palavras devem ser escolhidas a partir de uma lista mas deve também ter tido em conta o contexto da palavra na pergunta, podendo em alguns casos ser ignorada e não usada na interrogação ao SINO.

Outra forma de resolver o problema do sistema de recuperação de informação é deixar de o usar. Para isso deve ser resolvida a complexidade do sistema de forma a que seja capaz de gerar uma grande base de conhecimento, fazendo com que o sistema seja capaz de procurar as respostas em toda a colecção de documentos e não apenas num pequeno conjunto de documentos.

Para resolver o problema da ontologia vão continuar a ser estudadas novas formas de fazer a integração de várias ontologias de forma a obter uma boa ontologia genérica e que possa ser usada em qualquer contexto.

Uma outra forma para melhorar significativamente a qualidade da ontologia é criar relações entre as várias entidades que representem os sinónimos entre si. Com a criação destas relações torna-se possível a resolução de muitos problemas descritos na secção 4.4.2.

Para se conseguir resolver o processo de inferência na geração das respostas deve ser criado um mecanismo que consiga interpretar as anáforas dos documentos de texto e deve ser melhorado o módulo de geração das DRS's para que tenha um melhor comportamento com documentos de texto grandes.

Capítulo 5

Conclusões

Este trabalho tinha como principal objectivo a implementação dum protótipo dum sistema de perguntas e respostas para a Língua Portuguesa que conseguisse responder a algumas perguntas simples em Língua Portuguesa com respostas exactas.

Este objectivo foi claramente atingido visto que o sistema implementado conseguiu responder acertada-mente a 28.1% das perguntas feitas pelo QA@CLEF-2004, colocando-se em oitavo lugar na lista de sistemas que responderam a um maior número de perguntas com uma resposta correcta.

Estes resultados indicam que não só o sistema conseguiu alcançar os objectivos iniciais como também os superou.

Este sistema para além de ter conseguido satisfazer os objectivos iniciais conseguiu também mostrar que é possível transformar um protótipo de um sistema de perguntas e respostas numa aplicação que possa ser usada por qualquer utilizador.

Este sistema conseguiu obter bons resultados, no entanto não deixa de ter muitos problemas que limitam muito a capacidade do sistema responder a muitas perguntas.

Estes problemas devem-se principalmente à fraca escalabilidade que o sistema tem, à eficácia do motor de busca em obter os documentos relevantes para a pergunta e à qualidade da ontologia que foi usada.

Estes problemas podem ser resolvidos de várias de várias formas de modo a que se consiga obter um sistema mais robusto e que consiga responder a mais perguntas.

O problema do motor de busca está directamente ligada ao problema de escalabilidade de todo o sistema, ou seja, se o problema de escalabilidade do sistema for

resolvido deixa de ser necessário usar um sistema de recuperação de informação, ficando assim resolvido este problema.

A ontologia é também um grande problema do sistema, tentando-se no futuro fazer a construção duma ontologia genérica mais completa através de técnicas de combinações de várias ontologias mais pequenas.

Os bons resultados obtidos por este sistema fizeram com que a investigação de novas soluções e metodologias sejam estudadas de forma a serem integradas neste sistema com objectivo de se obterem melhores resultados.

Como trabalho futuro vai-se tentar resolver alguns dos problemas já descritos de forma a fazer uma evolução do sistema e fazer a integração de novas técnicas que ajudem a responder melhor a algumas perguntas como por exemplo a marcação de algumas anáforas de forma a que o sistema as consiga detectar e consiga responder a algumas perguntas tendo em conta as inúmeras anáforas existentes nos documentos de texto.

Actualmente este sistema de perguntas e respostas está integrado com um sistema de recuperação de informação, permitindo assim fazer perguntas em língua Portuguesa sobre as várias colecções de documentos existentes na aplicação.

Apêndice A

Sistemas participantes no QA@CLEF 2004

Aqui são descritos de uma forma muito geral alguns dos sistemas de perguntas e respostas que participaram no QA@CLEF 2004.

A.1 Question answering system for the French language

O objectivo deste projecto[22] foi a criação de um sistema de perguntas e respostas para a Língua Francesa e fazer a sua avaliação. Este sistema foi usado nas tarefas monolingue para a língua Francesa e multilingue, tendo como língua alvo a língua Francesa, e o Búlgaro, o Alemão, o Inglês, o Espanhol, o Italiano, o Holandês, e o Português como língua para as perguntas.

Este sistema aplica numa primeira fase um método clássico de recuperação de informação para extrair um pequeno número de parágrafos que possam conter informação que possa responder à pergunta. Depois dos parágrafos estarem extraídos, as perguntas e as frases dos parágrafos que foram recuperados pelo sistema de recuperação de informação são analisadas com um analisador sintáctico. Por fim, depois dos parágrafos relevantes terem sido recuperados e analisados, e as perguntas estarem analisadas é feita a extracção das respostas através de algumas técnicas de "matching", sendo escolhidas como respostas as frases que obtiverem um melhor "match".

Este sistema foi desenhado como sendo um sistema monolingue para a Língua Francesa, no entanto, este sistema participou nas tarefas multilingues, respondendo a uma pergunta com respostas em Francês. Para isso, as perguntas foram traduzidas da língua origem para o Francês através de vários sistemas de tradução

"on-line". Depois da pergunta estar traduzida para Francês, a a pergunta é passada ao sistema monolingue e é fornecida uma resposta.

Os resultados obtidos por este sistema mostram que a tradução das perguntas reduziu bastante a qualidade do sistema, pois os resultados das tarefas multilingue são piores que os resultados da tarefa monolingue. Na tarefa monolingue foi obtida uma percentagem de 24.5% de perguntas correctas, enquanto que na tarefa multilingue com melhores resultados(DE-FR) foi de 17%.

A.2 Cross-Language French-English Question Answering using the DLT System at CLEF 2004

Este sistema de perguntas e respostas[23] participa no QA@CLEF pela segunda vez, tentando nesta edição do CLEF fazer algumas melhorias ao sistema anterior.

A arquitectura base deste sistema de perguntas e respostas é o mais simples e normal possível, fazendo a identificação do tipo das perguntas, a análise e tradução das perguntas, a geração das interrogações para a recuperação de informação, a recuperação de documentos, a anotação dos textos e das perguntas, e por fim a extracção das perguntas.

A classificação das perguntas foi feita através do uso de combinações de simples palavras chave e padrões.

Análise das perguntas foi feita através da marcação da marcação "part-of-speech" das perguntas, sendo depois feita uma pesquisa superficial de vários tipos de frases. Depois da pergunta estar analisada, esta é traduzida através do uso de dois motores de tradução(Reverso e Wordlingo) e dum dicionário(Grand Dictionaire Terminolgique). Estas traduções eram depois combinadas de forma a obter uma única tradução. No caso de ser encontrada uma tradução no dicionário, as traduções feitas pelo motores de tradução eram ignoradas e usada a tradução do dicionário.

A geração das interrogações para a recuperação de informação era feita através das várias traduções encontradas para a pergunta, sendo criada uma interrogação booleana com várias alternativas através do uso de disjunções e conjunções.

A recuperação de informação foi feita através do motor de busca DTSearch com as interrogações previamente geradas.

A anotação dos textos foi feita através duma mistura de gramáticas e listas, havendo um total de 75 tipos.

A selecção da resposta é feita através do resultado da anotação dos textos e das perguntas, sendo escolhida uma instância da entidade que ocorre mais vezes

na pergunta traduzida e nos documentos que foram recuperados.

Este grupo apresentou dois "runs" que diferem apenas na forma como as perguntas foram traduzidas.

Este sistema conseguiu obter 19% de respostas correctas em 2004, enquanto que em 2003 apenas conseguiu 11.5% de respostas correctas, sendo assim feito uma boa melhoria no sistema.

A.3 Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question Answering System

Este sistema de perguntas e respostas[24] é a segunda vez que participa no QA@CLEF, no entanto, a versão usada na edição de 2004 é a versão de 2003 com muitas alterações. Este sistema foi usado no QA@CLEF 2004 na tarefa monolingue Alemã e na tarefa bilingue Alemão/Inglês.

O sistema usado na tarefa mono-lingue e na tarefa bilingue é praticamente o mesmo, havendo apenas algumas tarefas especificas para o sistema bilingue.

A arquitectura base do sistema consiste em:

- Motor linguístico
- Indexação da colecção de documentos
- Processador de perguntas em língua natural
- Sistema de recuperação de informação
- O processador das respostas

O motor linguístico está separado em duas partes: 1) um sistema que faz a anotação dos textos e delimita e faz a delimitação das frases com o uso de XML, e 2) um sistema que faz a análise sintáctica ás frases.

Para cada documento da colecção de documentos é feita uma indexação através do motor linguístico.

O processador de perguntas em língua natural, basicamente apenas determina o tipo de respostas que deve ser dada, quais as palavras que são mais relevantes

para a pergunta e o conjunto de entidades anotadas para ajudara na construção da resposta.

O sistema de recuperação de informação tenta encontrar frases na colecção de documentos que possa ser relevantes para a pergunta. Para isso é construída uma interrogação com base na análise de da pergunta em língua natural para depois ser feita a recuperação de informação.

O resultado do sistema de recuperação de informação é um conjunto de índices, onde cada índice aponta para uma única frase num documento anotado. Depois das frases relevantes terem sido obtidas, são escolhidas entidades que foram anotadas como sendo compatíveis com o tipo da resposta, sendo estas entidades possíveis respostas para a pergunta. Estas respostas são depois guardadas e é calculada frequência de cada uma das entidades, sendo escolhida como resposta a entidade que tiver uma maior frequência.

Este sistema gerou dois "runs", um para a tarefa monolingue Alemã e outro para a tarefa bilingue Alemão/Inglês, conseguindo obter resultados muito semelhantes para as duas tarefas, 25.3% de respostas certas para a tarefa monolingue e 23.3% para a tarefa bilingue.

Este sistema conseguiu obter melhores resultados em 2004 do que em 2003, pois as tarefas apresentadas no QA@CLEF 2004 eram mais complicadas que na edição de 2003, tanto na versão monolingue como na versão bilingue.

A.4 Cross-Language Question Answering at the University of Helsinki

Este sistema de perguntas e respostas[25], desenvolvido na Universidade de Helsinkia, foi desenvolvido para participar no QA@CLEF 2004 com o objectivo de responder a perguntas feitas em Finlandês usando uma colecção de documentos em Inglês. Uma particularidade deste sistema, é que foi desenhado para que possa ser configurado para usar outras linguagens. Este sistema de perguntas e respostas(*Tikka*) foi o primeiro sistemas de perguntas e resposta a usar o Finlandês.

Este sistema de perguntas e respostas está dividido em três módulos: módulo de processamento das perguntas, módulo de recuperação de informação e o módulo de extracção das respostas.

O módulo de processamento das perguntas faz numa primeira fase uma análise sintáctica à pergunta, numa segunda fase é determinado o tipo da pergunta e numa última fase faz a tradução da pergunta.

A determinação do tipo da pergunta é feita através do reconhecimento de algumas palavras na pergunta.

Depois da pergunta estar classificada, é passada a um sistema que faz a tradução da pergunta. Este sistema decide quais as palavras que devem ser traduzidas, como tratar os nomes próprios, os homónimos, as ambiguidades e as palavras que não aparecem no dicionário. Este sistema quais as palavras que devem ser usadas no sistema de recuperação de informação para obter a informação relevante para a pergunta. Estas decisões são feitas através duma árvore sintáctica da pergunta.

Depois do sistema de tradução ter encontrado quais as palavras que devem ser usadas no sistema de recuperação de informação, é construída uma interrogação que vai ser submetida ao motor de busca Managing Gigabytes que faz a recuperação dos documentos relevantes para a pergunta.

A primeira fase do módulo de extracção das respostas é a extracção de instâncias de padrões de respostas, usando para isso padrões de protótipos de respostas para cada tipo de pergunta. Estes padrões de protótipos de respostas são compostas por simples expressões regulares e por conjuntos de nomes próprios e palavras que foram previamente escolhidas a partir da pergunta.

Numa segunda fase, é escolhida uma das respostas candidatas como sendo a resposta que da pergunta. A resposta escolhida é a que aparecer com mais frequência na colecção de respostas candidatas.

Este sistema de perguntas e respostas conseguiu responder correctamente a 21 das perguntas 200, conseguindo assim uma percentagem de 10.88% de perguntas correctas. Destas 21 perguntas, 20 eram perguntas sobre factos e apenas uma das perguntas respondidas correctamente era uma pergunta de definição.

A.5 miraQA: Initial experiments in Question Answering

Este sistema de perguntas e respostas[26] foi desenvolvido para o QA@CLEF 2004, participando na tarefa Espanhola monolingue.

O miraQA é um sistema de perguntas e respostas que faz a extracção das respostas através de métodos estatísticos e usa o Google para fazer a recolha de dados para o treino do sistema.

Este sistema de perguntas e respostas foi implementado como sendo um sistema monolingue para a língua Espanhola, no entanto, foi desenhado para que possa

ser facilmente adaptado a outras línguas alvo. Para isso foram usados métodos estatísticos para fazer a extracção das respostas.

A arquitectura global deste sistema de perguntas e respostas segue uma estrutura clássica para os sistemas de perguntas e respostas, sendo constituído por um módulo que faz a análise das perguntas, um módulo que faz a recuperação da informação e por fim um módulo que faz a extracção das respostas.

O módulo que faz a análise das respostas faz a classificação das perguntas de acordo com uma taxonomia que foi criada manualmente e é composta por 17 classes. As perguntas são parcialmente analisadas, sendo depois usadas algumas regras simples para determinar o tipo da pergunta, o tipo da resposta e são adicionadas algumas "tags" semânticas a algumas partes da pergunta que possam estar relacionadas com a resposta.

O módulo de recuperação de informação faz a recuperação dos documentos mais relevantes para a pergunta e extrai as frases que contenham alguma das palavras da interrogação que foi usada no sistema de recuperação de informação. Esta interrogação é feita a partir das palavras da pergunta às quais foi adicionada informação semântica no módulo de análise das respostas. O sistema de recuperação de informação usado neste sistema é o Xapian, que atribui um valor de confiança a cada uma das frases obtidas pelo sistema de recuperação de informação.

O módulo de extracção das respostas usa um método de estatístico para fazer a extracção das respostas. Este método de extracção das respostas é feito em três etapas:

- Numa primeira fase, as frases obtidas pelo módulo de recuperação de informação são analisadas da mesma forma que as perguntas foram analisadas e é feito um treino do sistema com estas frases. Depois das frases estarem analisadas, as que têm termos da pergunta, estes são substituídos pela sua informação semântica. Por fim as frases são partidas em várias partes.

- Numa segunda fase é feito o reconhecimento das respostas através de um algoritmo que identifica a resposta como uma sequência de palavras geradas pelo tipo da resposta.

- Numa ultima etapa, as respostas são ordenadas através do valor de confiança do documento, da frase e da sequência de palavras encontrada para a pergunta.

Este sistema apresentou apenas um "run" para a tarefa monolingue Espanhola, fornecendo uma resposta exacta para cada uma das perguntas. Este sistema conseguiu responder correctamente a 18 perguntas num total de 200 perguntas. Estes

resultados são um pouco inferiores aos resultados obtidos por outros sistemas, podendo o mau desempenho do sistema ser atribuído ao facto deste sistema estar numa fase muito inicial.

A.6 Question Answering using Sentence Parsing and Semantic Nectwork Matching

O InSticht[27] é um sistema de perguntas e respostas para a língua Alemã. Este sistema participou no QA@CLEF 2004 na tarefa monolingue Alemã, conseguindo obter resultados bastante promissores.

Neste sistema, todos os documentos são analisados por um analisador sintáctico-semântico com o objectivo de representar cada uma das frases numa rede semântica.

Quando uma pergunta é submetida ao sistema, esta é analisada, sendo criada uma rede semântica para a pergunta e é calculado o tipo da pergunta. Depois da arvore semântica da pergunta estar construída, esta é expandida para outras redes semânticas equivalentes à inicial, usando para isso algumas regras de equivalência e variações de conceitos obtidas através de relações lexicais em várias fontes de conhecimento.

Na procura da resposta à pergunta, cada representação semântica que foi gerada para a pergunta é testada contra a representação semântica das frases dos documentos para ver se é possível fazer uma unificação entre alguma das representações da pergunta com a representação duma frase dos documentos.

No caso de haver uma unificação entre as duas representações é gerada uma resposta a partir da frase que conseguiu fazer a unificação com a pergunta através do uso de algumas regras de geração das respostas. Se houver mais do que uma resposta, então será escolhida a resposta mais longa e a que ocorrer mais vezes.

Este sistema consegui responder a 67 perguntas correctamente num total de 197 perguntas, pois existiam três perguntas com erros ortográficos.

Num teste não oficial, e depois deste sistema ter sido ligeiramente alterado ao nível da programação, este sistema conseguiu responder correctamente a 80 perguntas, mostrando assim o grande potencial deste sistema de perguntas e respostas.

Este sistema consegue ter uma grande precisão nas respostas(não nulas) fornecidas às perguntas, pois quando é fornecida uma resposta(não nula) á pergunta

está quase sempre correcta. No teste oficial do QA@CLEF apenas uma resposta não nula apresentada pelo sistema estava errada e no teste não oficial todas as respostas não nulas fornecidas estavam correctas.

A.7 First Evaluation of Esfinge - a question answering system for Portuguese

O sistema Esfinge[28] é um sistema de perguntas e respostas sobre domínios abertos na língua Portuguesa.

Este sistema de perguntas e respostas é baseado na arquitectura descrita por Eric Brill em (Brill, 2003). Esta arquitectura tenta poupar recursos no processamento das perguntas, e usar mais recursos para processar o grande volume de dados disponível na colecção de documento.

A arquitectura usada por este sistema é composta por quatro módulos: 1) Reformulação das perguntas, 2) Recuperação de informação, 3) Pesquisa das "N-grams", 4) Filtragem das "N-grams"

O módulo que faz a reformulação das perguntas tem como objectivo construir um conjunto de possíveis padrões para as possíveis respostas da pergunta. Por exemplo, para a pergunta "Em que ano chegou Vasco da Gama à Índia?", uma dos padrões criados seria "Vasco da Gama chegou á Índia em". Estes padrões foram feitos através de simples expressões regulares de Perl.

O módulo de recuperação de informação faz uma pesquisa dos padrões das respostas encontrados no módulo anterior na base de conhecimento, obtendo assim um conjunto de blocos de texto para cada pergunta. Estes blocos de texto são compostos pelo conjunto de 50 palavras que se encontram mais próximo do local onde foi encontrado o padrão no documento.

A pesquisa dos N-grams é feita em duas fases: 1) primeiro é calculada a distribuição dos N-grams das palavras(com tamanhos de 1 a 3) a partir dos blocos de texto que foram recuperados, e 2) por fim, a lista de N-grams das palavras é ordenado de acordo com a sua frequência, tamanho e relevância dos blocos de texto que deram origem aos N-grams.

Nesta fase o sistema tem uma lista ordenada de todas as respostas possíveis para a pergunta.

O módulo que faz a filtragem dos N-grams tem como objectivo excluir algumas das possíveis respostas através do uso de alguns filtros.

Um dos filtros serve para retirar respostas que estão contidas na própria pergunta. Por exemplo, na pergunta "Qual é a capital da Rússia?", a resposta "capital da Rússia" seria excluída do conjunto de possíveis respostas.

Outro filtro usado é um filtro que usa uma analisador morfológico as a morfologia de cada pergunta palavra da resposta. Este filtro classifica as palavras como sendo "interessantes" ou "não interessantes", sendo excluídas todas as respostas cuja primeira ou ultima palavra seja não interessante

Depois da filtragem das respostas é obtido um novo conjunto de respostas, sendo escolhida como resposta para a pergunta a resposta que tiver melhor pontuação.

Esta abordagem foi usada para criar um dos "runs",no entanto este sistema criou dois "runs" para submeter ao CLEF.

O segundo "run" submetido foi criado por uma variação deste sistema. Esta variação do sistema faz primeiro a pesquisa das respostas na Web com a ajuda do Google e do conjunto de padrões para as possíveis respostas. Os documentos recuperados pelo Google vão depois ser usados para extrair os blocos de texto e fazer todo o processamento das N-grams.

Por fim é adicionado um novo filtro que faz a pesquisa da resposta na colecção de documentos fornecida pelo CLEF com o objectivo de obter um documento que suporte a resposta encontrada na Web, sendo novamente extraída a resposta que tiver a maior pontuação.

Este sistema conseguiu obter 11% de respostas correctas no "run1" e 15.1% de respostas correctas no "run2"(com o uso da Web).No entanto, se forem consideradas as 5 primeiras respostas candidatas para cada pergunta estes valores sobem respectivamente para 16.1% e 18.6%, o que são resultados bastante interessantes.

A.8 TALP-QA System for Spanish at CLEF-2004

Este sistema de perguntas e resposta[29] tenta ser um sistema independente da língua, no entanto apenas participou na tarefa monolingue Espanhola no QA@CLEF 2004.

A arquitectura deste sistema de perguntas e respostas segue o esquema mais comum no desenho de sistemas de perguntas e respostas, estando dividido em

três partes, 1) processamento das perguntas, 2) recuperação de informação e 3) extracção da resposta.

O objectivo do processamento das perguntas neste sistema é fazer uma classificação das perguntas para descobrir qual o tipo de resposta que deve ser dada à pergunta e acrescentar informação às perguntas que vai ser usada nos módulos seguintes. Para o módulo de recuperação de informação é necessária a informação lexical e sintáctica e para o módulo de extracção da resposta é necessária a informação lexical, sintáctica e semântica, no entanto a informação mais importante é o tipo da pergunta.

O sistema de recuperação de informação cria e obtém blocos de texto relevantes para a pergunta através de um algoritmo iterativo. Para isso é usada a informação recolhida pelo sistema de processamento das perguntas. Estes blocos de texto são extraídos dos documentos que foram recuperados pelo motor de busca Lucene com a ajuda de uma interrogação criada a partir da pergunta, sendo cada um destes blocos de texto uma possível resposta para a pergunta. O algoritmo usado tenta encontrar um equilíbrio entre o número de blocos de texto encontrados, fazendo uma pesquisa mais geral se encontrar poucos blocos de texto e uma pesquisa mais especifica se encontrar muitos blocos de texto.

A extracção das respostas é feita de duas formas, sendo feita uma extracção para as resposta às perguntas sobre factos e um outro tipo de extracção sobre as respostas das perguntas de definição.

A extracção das respostas às perguntas sobre factos é feita em duas partes, sendo numa primeira parte extraídas as respostas candidatas e numa segunda parte a escolhida uma resposta das possíveis candidatas.

A extracção das respostas candidatas é feita sobre os blocos de texto que foram obtidos no sistema de recuperação de informação. Cada um destes blocos de texto é separado por frases, sendo depois atribuída uma pontuação a cada frase com base no seu conteúdo semântico que está de acordo com a pergunta. O conhecimento usado nesta fase é um conjunto de regras de extracção das respostas às quais está associado um valor de confiança. Para cada tipo de respostas usado está associado um conjunto de regras que serve para ajudar a fazer a extracção de respostas desse tipo

A selecção de uma resposta a partir do conjunto de respostas candidatas é feita com a ajuda do cálculo de algumas pontuações atribuídas a cada uma das respostas. Estas pontuações são calculada a através da pontuação atribuída à regra que foi usada para extrair a resposta, da pontuação que foi atribuída ao bloco de texto que contem a resposta e da pontuação semântica, sendo depois escolhida a resposta que obter uma melhor pontuação.

O objectivo das perguntas de definição no QA@CLEF 2004 é obter um segmento de texto da colecção de documentos que "define" uma pessoa ou uma organização. Estas definições aparecem normalmente como uma aposição, logo, as respostas são extraídas a partir das palavras que aparecem imediatamente antes ou depois da ocorrência do palavra que se pretende definir na pergunta. Outra forma de obter este tipo de respostas é através do uso de texto entre parêntesis que estejam perto do termo que se pretende definir.

A extracção deste tipo de perguntas foi feita em três etapas: 1) Análise das perguntas e extracção do termo que se pretende definir, 2) Cálculo da relevância das palavras para o termo que se pretende definir e 3) Selecção dos fragmentos de texto com mais informação.

A análise das perguntas e extracção do termo que se pretende definir é feito através do mesmo módulo que faz a análise às perguntas sobre factos, devolvendo o termo que se pretende que se pretende definir e qual o seu tipo. A identificação deste tipo serve para escolher heurísticas mais especificas que ajudam na extracção da pergunta.

O cálculo da relevância das palavras mede a forma como as palavras estão relacionadas com o termo que se pretende definir. Para calcular este valor é feita uma janela de texto que contem as 15 palavras antes e as 15 palavras depois da ocorrência do termo que se pretende definir, sendo depois calculado o número de ocorrências das palavras que pertencem a essa janela, sendo as palavras mais relevantes as que aparecerem menos vezes.

A selecção dos blocos de texto mais relevantes é feita através do cálculo da densidade de informação que cada bloco de texto, sendo esta densidade a soma do valor de relevância de cada palavra do bloco de texto. Depois de calculado este valor para todas as respostas, é escolhida a resposta que tiver uma maior densidade de informação.

Este sistema conseguiu fornecer uma resposta para cada uma das 200 perguntas do conjunto de teste da tarefa Espanhola monolingue, conseguindo responder correctamente a 48 respostas num "run" e a 52 noutro "run", conseguindo um total de 25% e 26% de respostas correctas.

Apêndice B

Perguntas e respostas no QA@CLEFF-2004

Neste apêndice estão todas as perguntas feitas pelo QA@CLEF-2004, bem como as respostas que dadas a cada uma da perguntas pelo nosso sistema.

B.1 Perguntas do QA@CLEF-2004

Neste secção pode-se ver a lista de todas as perguntas feitas pelo QA@CLEF-2004 para a língua Portuguesa, estando estas representadas num formato definido pelo CLEF.

```
F PT PT 0001 Em que cidade se encontra a
prisão de San Vittore?
F PT PT 0002 Onde era o campo de concentração
de Auschwitz?
F PT PT 0003 Quem foi o autor de "Mein Kampf"?
F PT PT 0004 Qual é a capital da Rússia?
F PT PT 0005 Quem foi o primeiro presidente dos
Estados Unidos?
F PT PT 0006 Como morreu Jimi Hendrix?
F PT PT 0007 Com quem se casou Michael Jackson?
F PT PT 0008 Em que género musical se distingue
Michael Jackson?
D PT PT 0009 O que é a Mossad?
F PT PT 0010 Quantos crimes são atribuídos ao
Monstro de Florença?
```

F PT PT 0011 Quantos desempregados há na Europa?
F PT PT 0012 Quantas religiões monoteístas
há no mundo?
F PT PT 0013 Quantos judeus existem no mundo?
F PT PT 0014 Quantos detidos há no Corredor
da Morte na Califórnia?
D PT PT 0015 O que é a UNICEF?
F PT PT 0016 Nomeie uma pessoa acusada de
pedofilia.
D PT PT 0017 Quem é Jean-Bertrand Aristide?
F PT PT 0018 Mencione um cetáceo.
F PT PT 0019 Quem escreveu "Ulisses"?
F PT PT 0020 Onde se situa o CERN?
D PT PT 0021 Quem é Yves Saint-Laurent?
F PT PT 0022 Em que dia calha o solstício de
verão?
F PT PT 0023 Onde fica o Museu do Hermitage?
F PT PT 0024 De que são feitos os cabos de
fibra óptica?
F PT PT 0025 Que forma de governo tem a França?
F PT PT 0026 Qual o nome da mulher de Kurt Cobain?
F PT PT 0027 Qual o vulcão activo mais alto da Europa?
F PT PT 0028 O que significa "Forza Italia"?
F PT PT 0029 Quando foi lançada a sonda
espacial Ulisses?
D PT PT 0030 O que é a maçonaria?
F PT PT 0031 Quem foi o último czar da Rússia?
F PT PT 0032 Qual o acrónimo da Amnistia Internacional?
F PT PT 0033 Onde se entregam os Óscares?
F PT PT 0034 Onde fica o arquipélago de Svalbard?
F PT PT 0035 Onde se realizou a Conferência
Mundial da Mulher?
F PT PT 0036 Indique uma companhia de fast-food.
D PT PT 0037 Quem foi Rosa Chacel?
D PT PT 0038 Quem é Christo?
D PT PT 0039 O que são as FARC?
F PT PT 0040 Qual a abreviatura do Exército Popular
de Libertação do Sudão?
F PT PT 0041 Em que ano é que o "War Powers Act" foi
aprovado pelo Congresso americano?

F PT PT 0042 Esmirna fica em que país?
F PT PT 0043 Qual a localização de Tipaza?
F PT PT 0044 Quem é o fundador da Motown?
F PT PT 0045 Como se chama a filha do líder chinês
Deng Xiaoping?
D PT PT 0046 O que é o FSK?
F PT PT 0047 Em que país fica Vukovar?
F PT PT 0048 Em que cidade americana se encontra
o Museu Warhol?
D PT PT 0049 Quem é Andy Warhol?
F PT PT 0050 Quem descobriu o vírus da sida?
F PT PT 0051 Quem é a ministra do Ambiente alemã?
F PT PT 0052 Mencione um bonecreiro.
F PT PT 0053 Quem é o presidente da UEFA?
D PT PT 0054 O que é a MTV?
F PT PT 0055 Quem é o líder do KwaZulu?
D PT PT 0056 O que é a NASA?
F PT PT 0057 Quem planeou o Palácio dos Desportos
São Jorge em Barcelona?
F PT PT 0058 Em que equipa de básquete joga Shaquille
O'Neill?
F PT PT 0059 Quem é o presidente da Câmara dos
Representantes americana?
F PT PT 0060 Em que ano foi assassinado o presidente
chileno Salvador Allende?
D PT PT 0061 Quem é Marvin Minsky?
F PT PT 0062 Como se intitula a autobiografia de
Nelson Mandela?
F PT PT 0063 Como se chama a viúva do falecido
presidente de Moçambique, Samora Machel?
D PT PT 0064 Quem é João Havelange?
F PT PT 0065 O que significa a abreviatura OUA?
F PT PT 0066 Onde fica Hyde Park?
F PT PT 0067 Que significa a abreviatura AWACS
nos aviões AWACS?
F PT PT 0068 Onde está preso Hugo Lacour?
F PT PT 0069 Quantos assinantes tem a MSN?
D PT PT 0070 O que é a UNICE?
F PT PT 0071 Quem foi forçado a demitir-se de
governador da Caríntia em 1991?

F PT PT 0072 Qual é a maior empresa industrial
da Finlândia?
F PT PT 0073 Qual o lucro do grupo electrónico
e de telecomunicações finlandês Nokia em 1994?
D PT PT 0074 O que é o CERN?
F PT PT 0075 Quantos estados-membros tem o CERN?
D PT PT 0076 Quem é Kevin Mitnick?
F PT PT 0077 Quando foi criado o CERN?
F PT PT 0078 Que produz a MCC?
F PT PT 0079 Qual o monte mais alto do mundo?
F PT PT 0080 Onde fica Halifax?
D PT PT 0081 Quem é Umberto Bossi?
F PT PT 0082 Onde fica o La Scala?
F PT PT 0083 Onde fica a sede da UNESCO?
F PT PT 0084 Quem é o realizador de "Nikita"?
F PT PT 0085 Quantos anos de residência são
necessários para obter a nacionalidade suíça?
D PT PT 0086 O que é o GIA?
F PT PT 0087 Qual o cargo de Redha Malek em 1994?
F PT PT 0088 Quem foi eleito presidente do Conselho
Geral da Guiana?
F PT PT 0089 Que quantia exige o FC Sevilha de Diego
Maradona?
F PT PT 0090 Como se chama o ministro das Finanças polaco?
F PT PT 0091 Como morreu Juvénal Habyarimana?
F PT PT 0092 Quem foi derrotado por Andrei Medveded na
final do Torneio de Monte-Carlo?
F PT PT 0093 Qual a nacionalidade do tenista Sergi Bruguera?
F PT PT 0094 Qual a superfície da República da Chechénia?
F PT PT 0095 Qual o nome do partido político de Ntsu Mokhele
, primeiro-ministro do Lesoto?
F PT PT 0096 Qual o nome do primeiro-ministro do Ruanda?
F PT PT 0097 De que país é a escritora Taslima Nasreen?
F PT PT 0098 Qual o cargo de Albert Reynolds na Irlanda?
F PT PT 0099 Qual a taxa de desemprego nos Estados Unidos
no final de 1994?
F PT PT 0100 Quando tiveram lugar as eleições
europeias de 1994?
F PT PT 0101 Onde fica a Esfinge de Gizé?
F PT PT 0102 Onde é Izhevsk?

F PT PT 0103 Onde fica o Estádio José Alvalade?
F PT PT 0104 Onde vive José Saramago?
F PT PT 0105 Onde nasceu Nelson Mandela?
F PT PT 0106 Qual o maior satélite de Júpiter?
F PT PT 0107 Onde fica Turku?
F PT PT 0108 Qual a antiga capital da Polónia?
F PT PT 0109 Em que distrito fica Paredes de Coura?
F PT PT 0110 Onde fica Sosnovy Bor?
F PT PT 0111 Em que ilha fica Ponta Delgada?
F PT PT 0112 Onde é que nasceu Álvaro Cunhal?
F PT PT 0113 Onde é o hospital Júlio de Matos?
F PT PT 0114 Qual é o país mais pequeno da União Europeia?
F PT PT 0115 Onde fica Gabrovo?
F PT PT 0116 Qual o estado mais setentrional dos EUA?
F PT PT 0117 Qual é a capital da Bielorússia?
F PT PT 0118 Onde desagua o rio Cubango?
F PT PT 0119 Em que cidade o Mosela encontra o Reno?
F PT PT 0120 Em que estado do Brasil fica Campo Grande?
F PT PT 0121 Onde se situa Tianjin?
F PT PT 0122 Onde é a Ilha do Diabo?
F PT PT 0123 Quem inventou o saxofone?
F PT PT 0124 Quem escreveu "O Principezinho"?
F PT PT 0125 Quem é o recordista mundial do salto à vara?
F PT PT 0126 Quem é a "diva dos pés descalços"?
F PT PT 0127 Quem é o secretário-geral do PCP?
F PT PT 0128 De quem é filha Martine Aubry?
F PT PT 0129 Quem é o Presidente da Câmara de Lisboa?
F PT PT 0130 Quem é o Presidente da Câmara de Lamego?
F PT PT 0131 Quem é o embaixador de Portugal em França?
F PT PT 0132 Com quem casou Whoppi Goldberg?
F PT PT 0133 Quem foi o primeiro presidente dos
Estados Unidos?
F PT PT 0134 Quem é o ministro-presidente da
Renânia-Palatinado?
F PT PT 0135 Quem foi o último governador de Timor Leste?
F PT PT 0136 Quem era o marido de Vieira da Silva?
F PT PT 0137 Quem é o capitão do FC Porto?
F PT PT 0138 Quem é o imã da mesquita de Lisboa?
F PT PT 0139 Quem realizou o filme "Lisbon Story"?
F PT PT 0140 Quem é a ministra sueca do ambiente?

F PT PT 0141 Como se chama a rainha da Dinamarca?
F PT PT 0142 Quem é o padroeiro de Penafiel?
F PT PT 0143 Que grupo matou Aldo Moro?
F PT PT 0144 De que grupo é vocalista Teresa Salgueiro?
F PT PT 0145 Que equipa venceu a Taça CERS em
hóquei em patins?
F PT PT 0146 De que clube é treinador Bobby Robson?
F PT PT 0147 Que empresa tem uma refinaria
em Leça da Palmeira?
F PT PT 0148 A que partido pertence Duarte Lima?
F PT PT 0149 Quem financia as IPSS?
F PT PT 0150 Quantos submarinos tem a marinha portuguesa?
F PT PT 0151 Quantos municípios há em Portugal?
F PT PT 0152 Qual o comprimento da Ponte do Freixo?
F PT PT 0153 Quantos anos tem Inês de Medeiros?
F PT PT 0154 Qual a distância de Braga a Guimarães?
F PT PT 0155 Qual a altura do K2?
F PT PT 0156 Qual o valor da dívida da Eurotunnel?
F PT PT 0157 Qual a área da Baixa-Saxónia?
F PT PT 0158 Quantos habitantes tem a República Dominicana?
F PT PT 0159 Quantos golos marcou Eusébio na sua carreira?
F PT PT 0160 A que velocidade viaja a luz?
F PT PT 0161 Quando foram criadas as FPLM (Forças Populares
de Libertação de Moçambique)?
F PT PT 0162 Quando foi a independência de Cabo Verde?
F PT PT 0163 Quando estreia o filme "Lisbon Story"?
F PT PT 0164 Quando foi aprovada a Declaração Universal
dos Direitos do Homem?
F PT PT 0165 Quando morreu Salvador Allende?
F PT PT 0166 Quando morreu Simão Bolívar?
F PT PT 0167 Em que dia se comemora a independência do Brasil?
F PT PT 0168 Quando se tornou "A Portuguesa" hino nacional?
F PT PT 0169 Em que ano ocorreu o 25 de Abril?
F PT PT 0170 Em que embateu o Titanic?
F PT PT 0171 Qual o símbolo de liderança da Volta a Itália?
F PT PT 0172 O que foi erguido em 13 de Agosto de 1961?
F PT PT 0173 A que era alérgico Mel Blanc?
F PT PT 0174 Que país é campeão do mundo de futebol?
F PT PT 0175 Como morreu Pasolini?
F PT PT 0176 Como se tornou o Brasil tetracampeão mundial

```
de futebol?
F PT PT 0177 Qual foi o primeiro filme sonoro português?
F PT PT 0178 Que vende Fausto ao Diabo?
F PT PT 0179 Que animal é o símbolo da Namíbia?
F PT PT 0180 Qual o pseudónimo de Álvaro Cunhal?
F PT PT 0181 Qual é a nacionalidade de Yordan Letchkov?
F PT PT 0182 Qual a nacionalidade de Hercule Poirot?
F PT PT 0183 O que era Napoleão III a Napoleão Bonaparte?
F PT PT 0184 De que material são os frisos do Parténon?
F PT PT 0185 Qual a patente de Alfred Dreyfus?
F PT PT 0186 De que cor é a neve?
F PT PT 0187 Qual é a moeda iraquiana?
F PT PT 0188 Qual o endereço da Livraria Barata?
D PT PT 0189 Quem é Leonor Beleza?
D PT PT 0190 Quem é Arnold Ruutel?
D PT PT 0191 Quem é Wim Duisenberg?
D PT PT 0192 Quem é Rocha Vieira?
D PT PT 0193 Quem é Guilherme da Fonseca?
D PT PT 0194 Quem é Fernando Gomes?
D PT PT 0195 Quem é Valentina Terechkova?
D PT PT 0196 Quem é Jorge Amado?
D PT PT 0197 O que é o PC do B?
D PT PT 0198 O que é o PSN?
D PT PT 0199 O que é o CSKA?
D PT PT 0200 O que é a Vigor?
```

B.2 Respostas para o QA@CLEF 2004

Nesta secção estão todas as respostas dadas pelo nosso sistema a cada uma das perguntas feitas pelo QA@CLEF-2004, estando representadas no formato definido pelo próprio CLEF.

```
F 0001 PTUE041ptpt 1 PUBLICO-19940724-039 Milão
F 0002 PTUE041ptpt 1 PUBLICO-19940804-089 sul Polónia
F 0003 PTUE041ptpt 1 PUBLICO-19940828-042 Hitler
F 0004 PTUE041ptpt 1 PUBLICO-19940430-114 Moscovo
F 0005 PTUE041ptpt 1  NIL
F 0006 PTUE041ptpt 1 PUBLICO-19940310-008 asfixia
F 0007 PTUE041ptpt 1 PUBLICO-19940803-003 Lisa Marie Presley
```

```
F 0008 PTUE041ptpt 1  NIL
D 0009 PTUE041ptpt 1 PUBLICO-19940522-041 serviço secreto
israelita
F 0010 PTUE041ptpt 1  NIL
F 0011 PTUE041ptpt 1 PUBLICO-19940211-102 20 milhão
F 0012 PTUE041ptpt 1 PUBLICO-19951127-121 três
F 0013 PTUE041ptpt 1 PUBLICO-19950508-053 800 mil 40
F 0014 PTUE041ptpt 1  NIL
D 0015 PTUE041ptpt 1 PUBLICO-19940103-006 organização
Nações Unidas
F 0016 PTUE041ptpt 1 PUBLICO-19950802-135 Viena
D 0017 PTUE041ptpt 1 PUBLICO-19940307-043 Haiti
F 0018 PTUE041ptpt 1 PUBLICO-19940106-066 -
F 0019 PTUE041ptpt 1 PUBLICO-19940829-004 Xião Qian
F 0020 PTUE041ptpt 1 PUBLICO-19940427-037 Genebra
D 0021 PTUE041ptpt 1 PUBLICO-19951124-145 costureiro como
Sonya Rykel Christian Dior Nina Ricci Hermés Azzaro Cacharel
Guy Laroche Balenciaga Paco Rabanne Ted Lapidus Pierre Balmain
F 0022 PTUE041ptpt 1  NIL
F 0023 PTUE041ptpt 1  NIL
F 0024 PTUE041ptpt 1  NIL
F 0025 PTUE041ptpt 1  NIL
F 0026 PTUE041ptpt 1  NIL
F 0027 PTUE041ptpt 1  NIL
F 0028 PTUE041ptpt 1  NIL
F 0029 PTUE041ptpt 1  NIL
D 0030 PTUE041ptpt 1  NIL
F 0031 PTUE041ptpt 1  NIL
F 0032 PTUE041ptpt 1  NIL
F 0033 PTUE041ptpt 1  NIL
F 0034 PTUE041ptpt 1  NIL
F 0035 PTUE041ptpt 1  NIL
F 0036 PTUE041ptpt 1 PUBLICO-19951122-073 McDonald' s
D 0037 PTUE041ptpt 1 PUBLICO-19940731-057 Madrid escritora
espanhol ano
D 0038 PTUE041ptpt 1 PUBLICO-19940125-094 Alberto
Barceló Klein
D 0039 PTUE041ptpt 1 PUBLICO-19940124-089 presumível
guerrilheiro Forças Armadas Revolucionárias marxista
F 0040 PTUE041ptpt 1  NIL
```

```
F 0041 PTUE041ptpt 1  NIL
F 0042 PTUE041ptpt 1  NIL
F 0043 PTUE041ptpt 1  NIL
F 0044 PTUE041ptpt 1  NIL
F 0045 PTUE041ptpt 1  NIL
D 0046 PTUE041ptpt 1 PUBLICO-19940616-048 contra-informação
F 0047 PTUE041ptpt 1  NIL
F 0048 PTUE041ptpt 1  NIL
D 0049 PTUE041ptpt 1 PUBLICO-19940417-005 artista contemporâneo
F 0050 PTUE041ptpt 1 PUBLICO-19940401-053 Luc Montagnier
F 0051 PTUE041ptpt 1 PUBLICO-19950314-046 ministra Angela Merkel
F 0052 PTUE041ptpt 1 PUBLICO-19950505-014 -
F 0053 PTUE041ptpt 1 PUBLICO-19940219-013 Lennart Johansson
presidente UEFA um
D 0054 PTUE041ptpt 1 PUBLICO-19940221-030 estação norte-americana
 provavelmente outro grande operador
F 0055 PTUE041ptpt 1  NIL
D 0056 PTUE041ptpt 1 PUBLICO-19940202-043 Guerra das Estrelas
 agência espacial americano
F 0057 PTUE041ptpt 1  NIL
F 0058 PTUE041ptpt 1  NIL
F 0059 PTUE041ptpt 1  NIL
F 0060 PTUE041ptpt 1  NIL
D 0061 PTUE041ptpt 1 PUBLICO-19940423-049 todo o país
F 0062 PTUE041ptpt 1  NIL
F 0063 PTUE041ptpt 1  NIL
D 0064 PTUE041ptpt 1 PUBLICO-19940107-009 brasileiro presidente FIFA
F 0065 PTUE041ptpt 1  NIL
F 0066 PTUE041ptpt 1 PUBLICO-19940719-117 Londres
F 0067 PTUE041ptpt 1  NIL
F 0068 PTUE041ptpt 1  NIL
F 0069 PTUE041ptpt 1  NIL
D 0070 PTUE041ptpt 1 PUBLICO-19940513-124 relatório ainda
interino União das Confederações da Indústria
Empregadores da Europa Público
F 0071 PTUE041ptpt 1  NIL
F 0072 PTUE041ptpt 1  NIL
F 0073 PTUE041ptpt 1  NIL
D 0074 PTUE041ptpt 1 PUBLICO-19940422-045 Laboratório
Europeu de Física das Partículas enquanto observação
```

```
F 0075 PTUE041ptpt 1  NIL
D 0076 PTUE041ptpt 1 PUBLICO-19950311-160 um
F 0077 PTUE041ptpt 1 PUBLICO-19940913-051 1954 19
F 0078 PTUE041ptpt 1  NIL
F 0079 PTUE041ptpt 1 PUBLICO-19940901-080 Evereste
F 0080 PTUE041ptpt 1 PUBLICO-19941009-021 Canadá
D 0081 PTUE041ptpt 1 PUBLICO-19940107-049 líder Liga Norte
F 0082 PTUE041ptpt 1  NIL
F 0083 PTUE041ptpt 1  NIL
F 0084 PTUE041ptpt 1  NIL
F 0085 PTUE041ptpt 1  NIL
D 0086 PTUE041ptpt 1 PUBLICO-19940227-048 Grupo Islâmico Armado
F 0087 PTUE041ptpt 1  NIL
F 0088 PTUE041ptpt 1  NIL
F 0089 PTUE041ptpt 1  NIL
F 0090 PTUE041ptpt 1  NIL
F 0091 PTUE041ptpt 1  NIL
F 0092 PTUE041ptpt 1  NIL
F 0093 PTUE041ptpt 1  NIL
F 0094 PTUE041ptpt 1  NIL
F 0095 PTUE041ptpt 1  NIL
F 0096 PTUE041ptpt 1  NIL
F 0097 PTUE041ptpt 1  NIL
F 0098 PTUE041ptpt 1  NIL
F 0099 PTUE041ptpt 1  NIL
F 0100 PTUE041ptpt 1  NIL
F 0101 PTUE041ptpt 1  NIL
F 0102 PTUE041ptpt 1  NIL
F 0103 PTUE041ptpt 1 PUBLICO-19950529-025 Lisboa
F 0104 PTUE041ptpt 1 PUBLICO-19950417-094 Lanzarote
F 0105 PTUE041ptpt 1  NIL
F 0106 PTUE041ptpt 1 PUBLICO-19951022-048 Ganimedes
F 0107 PTUE041ptpt 1 PUBLICO-19950928-111 Finlândia
F 0108 PTUE041ptpt 1  NIL
F 0109 PTUE041ptpt 1  NIL
F 0110 PTUE041ptpt 1 PUBLICO-19950601-125 São Petersburgo
F 0111 PTUE041ptpt 1  NIL
F 0112 PTUE041ptpt 1  NIL
F 0113 PTUE041ptpt 1 PUBLICO-19940130-123 Lisboa
F 0114 PTUE041ptpt 1  NIL
```

```
F 0115 PTUE041ptpt 1 PUBLICO-19940910-118 Bulgária
F 0116 PTUE041ptpt 1 PUBLICO-19950205-091 Alasca
F 0117 PTUE041ptpt 1  NIL
F 0118 PTUE041ptpt 1  NIL
F 0119 PTUE041ptpt 1 PUBLICO-19950128-158 Coblença
F 0120 PTUE041ptpt 1  NIL
F 0121 PTUE041ptpt 1 PUBLICO-19950227-001 China
F 0122 PTUE041ptpt 1  NIL
F 0123 PTUE041ptpt 1  NIL
F 0124 PTUE041ptpt 1  NIL
F 0125 PTUE041ptpt 1 PUBLICO-19950615-016 Bubka
O recordista mundial
F 0126 PTUE041ptpt 1  NIL
F 0127 PTUE041ptpt 1 PUBLICO-19940305-183 Carvalhas
F 0128 PTUE041ptpt 1  NIL
F 0129 PTUE041ptpt 1  NIL
F 0130 PTUE041ptpt 1  NIL
F 0131 PTUE041ptpt 1 PUBLICO-19940317-169 França
F 0132 PTUE041ptpt 1  NIL
F 0133 PTUE041ptpt 1  NIL
F 0134 PTUE041ptpt 1 PUBLICO-19940127-068 Kohl Joe Klein
F 0135 PTUE041ptpt 1 PUBLICO-19940329-112 Mário Lemos Pires
F 0136 PTUE041ptpt 1  NIL
F 0137 PTUE041ptpt 1 PUBLICO-19941003-024 FC
F 0138 PTUE041ptpt 1  NIL
F 0139 PTUE041ptpt 1  NIL
F 0140 PTUE041ptpt 1 PUBLICO-19950610-153 Anna Lindh
F 0141 PTUE041ptpt 1  NIL
F 0142 PTUE041ptpt 1 PUBLICO-19950823-073 lágrima
olho Depois do S. Martinho padroeiro Penafiel este cidade amanhã
F 0143 PTUE041ptpt 1  NIL
F 0144 PTUE041ptpt 1  NIL
F 0145 PTUE041ptpt 1  NIL
F 0146 PTUE041ptpt 1  NIL
F 0147 PTUE041ptpt 1  NIL
F 0148 PTUE041ptpt 1  NIL
F 0149 PTUE041ptpt 1 PUBLICO-19950626-066 Segurança Social
F 0150 PTUE041ptpt 1  NIL
F 0151 PTUE041ptpt 1  NIL
F 0152 PTUE041ptpt 1  NIL
```

```
F 0153 PTUE041ptpt 1  NIL
F 0154 PTUE041ptpt 1  NIL
F 0155 PTUE041ptpt 1  NIL
F 0156 PTUE041ptpt 1  NIL
F 0157 PTUE041ptpt 1  NIL
F 0158 PTUE041ptpt 1  NIL
F 0159 PTUE041ptpt 1  NIL
F 0160 PTUE041ptpt 1  NIL
F 0161 PTUE041ptpt 1 PUBLICO-19940817-054 -
F 0162 PTUE041ptpt 1  NIL
F 0163 PTUE041ptpt 1  NIL
F 0164 PTUE041ptpt 1  NIL
F 0165 PTUE041ptpt 1  NIL
F 0166 PTUE041ptpt 1 PUBLICO-19950131-056 17 dezembro 1830
F 0167 PTUE041ptpt 1  NIL
F 0168 PTUE041ptpt 1  NIL
F 0169 PTUE041ptpt 1  NIL
F 0170 PTUE041ptpt 1  NIL
F 0171 PTUE041ptpt 1  NIL
F 0172 PTUE041ptpt 1  NIL
F 0173 PTUE041ptpt 1  NIL
F 0174 PTUE041ptpt 1  NIL
F 0175 PTUE041ptpt 1  NIL
F 0176 PTUE041ptpt 1  NIL
F 0177 PTUE041ptpt 1  NIL
F 0178 PTUE041ptpt 1  NIL
F 0179 PTUE041ptpt 1  NIL
F 0180 PTUE041ptpt 1 PUBLICO-19950802-006 José Fonseca
F 0181 PTUE041ptpt 1  NIL
F 0182 PTUE041ptpt 1 PUBLICO-19940327-069 Hercule Poirot
F 0183 PTUE041ptpt 1  NIL
F 0184 PTUE041ptpt 1  NIL
F 0185 PTUE041ptpt 1  NIL
F 0186 PTUE041ptpt 1  NIL
F 0187 PTUE041ptpt 1  NIL
F 0188 PTUE041ptpt 1  NIL
D 0189 PTUE041ptpt 1 PUBLICO-19940120-090 presidente
Comissão Parlamentar dos Assuntos Europeus recentemente em público
D 0190 PTUE041ptpt 1 PUBLICO-19950307-057 Soviete Supremo Estónia
D 0191 PTUE041ptpt 1  NIL
```

```
D 0192 PTUE041ptpt 1 PUBLICO-19940118-098 Macau
D 0193 PTUE041ptpt 1 PUBLICO-19940904-066 inconstitucionalidade
juiz-conselheiro Antero Alves Monteiro Dinis Maria Fernanda dos
 Santos Martins da Palma Pereira Luís Nunes de Almeida Alberto
Tavares da Costa Armindo Ribeiro Mendes Maria da Assunção Esteves
 José Manuel Cardoso da Costa declaração
D 0194 PTUE041ptpt 1 PUBLICO-19940119-156 Presidente da Câmara
 Municipal do Porto
D 0195 PTUE041ptpt 1 PUBLICO-19950630-049 16.6.63 soviético primeiro
 mulher cosmonauta
D 0196 PTUE041ptpt 1 PUBLICO-19940325-093 escritor
D 0197 PTUE041ptpt 1  NIL
D 0198 PTUE041ptpt 1 PUBLICO-19940401-115 O Partido da Solidariedade
Nacional
D 0199 PTUE041ptpt 1 PUBLICO-19941011-020 Moscovo
D 0200 PTUE041ptpt 1  NIL
```

Bibliografia

[1] J. Parikh and M. Narasimha Murty. Adapting question answering techniques to the web. 2002.

[2] Diego Mollá, Rolf Schwitter, Fabio Rinaldi, Jmes Dowdall, and Michael Hess. Nlp for answer extraction in technical domains.

[3] Raymond J. Mooney. Learning for semantic interpretation: Scaling up without dumbling down. In *Workshop Notes for the Workshop on Learning in Logic, Bled, Slovenia*, June 1999.

[4] Cynthia A. Thompson and Raymond J. Mooney. Semantic lexicon acquisition for learnig parsers. 1997.

[5] Nick Trebon. Natural language processing.

[6] P.J.Hancox. A brief history of natural language processing.

[7] Patrick Doyle. Natural language ai qual summary.

[8] Sanda M. Harabagiu, Marius A. Pasca, and Steven J. Maiorano. Experiments with open-domain textual question answering.

[9] Boris Katz, Jimmy Lin, Daniel Loreto, Wesley Hildebrandt, Matthew Bilotti, Sue Felshin, Aaron Fernandes, Gregory Marton, and Federico Mora. Integrating web-based and corpus-based techniques for question answering.

[10] Jochen L. Leidner, Johan Bos, Tiphaine Dalmas ann James R. Curran, Stephen Clark, Colin J. Bannard, Bonnie Webber, and Mark Steedman. Qed: The edinburgh trec-2003 question answering system. 2003.

[11] Hans Kamp and Uwe Reyle. *From Discourse to Logic:An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Dordrecht: D. Reidel, 1993.

[12] Paulo Quaresma, Luís Quintano, Irene Rodrigues, José Saias, and Pedro Salgueiro. The university of Évora approch to qa@clef-2004. In *Cross Language Evaluation Forum, Results of the CLEF 2004 CROSS-Language SYstem Evaluation Campaign*, Universidade de Évora, Portugal, 2004. Departamento de Informática, Universidade de Évora.

[13] José Saias. Uma metodologia para a construção automática de ontologias e a sua aplicação em sistemas de recuperação de informação – a methodology for the automatic creation of ontologies and its application in information retrieval systems. Master's thesis, University of Évora, Portugal, 2003. In Portuguese.

[14] José Saias and Paulo Quaresma. Using nlp techniques to create legal ontologies in a logic programming based web information retrieval system. In *Workshop on Legal Ontologies and Web based legal information management of the 9th International Conference on Artificial Intelligence and Law*, Edinburgh, Scotland, June 2003.

[15] Salvador Abreu, Paulo Quaresma, Luis Quintano, and Irene Rodrigues. A dialogue manager for accessing databases. In *13th European-Japanese Conference on Information Modelling and Knowledge Bases*, pages 213–224, Kitakyushu, Japan, June 2003. Kyushu Institute of Technology. To be published by IOS Press.

[16] Paulo Quaresma and Irene Pimenta Rodrigues. A natural language interface for information retrieval on semantic web documents. In E. Menasalvas, J. Segovia, and P. Szczepaniak, editors, *AWIC'2003 - Atlantic Web Intelligence Conference*, Lecture Notes in Artificial Intelligence LNCS/LNAI 2663, pages 142–154, Madrid, Spain, May 2003. Springer-Verlag.

[17] Salvador Abreu. Isco: A practical language for heterogeneous information system construction. In *Proceedings of INAP'01*, Tokyo, Japan, October 2001. INAP.

[18] Paulo Quaresma and Irene Pimenta Rodrigues. PGR: Portuguese atorney general's office decisions on the web. In Osamu Yoshie, editor, *Proceedings of the 14th International Conference on Applications of Prolog*, University of Tokyo, Tokyo, Japan, October 2001. REN Associates, Inc. ISSN 1345-0980. To be published by Springer Verlag's LNAI.

[19] G. Greenleaf, A. Mowbray, and G. King. Law on the net via austlii - 14 m hypertext links can't be right? In *In Information Online and On Disk'97 Conference, Sydney*, 1997.

[20] Eckhard Bick. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus University Press, 2000.

[21] Jerry Hobs, Mark Stickel, Douglas Applet, and Paul Martin. Interpretation as abduction.

[22] Laura Perret. Question answering system for the french language. In *Cross Language Evaluation Forum, Results of the CLEF 2004 CROSS-Language SYstem Evaluation Campaign*, University of Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland, 2004. Institute interfacultaire d'informatique.

[23] Richard F. E. Sutclife, Igal Gabbay, Michael Mulchy, and Aoife O'Gorman. Cross-language french-english question answering using the dlt system ad clef 2004. In *Cross Language Evaluation Forum, Results of the CLEF 2004 CROSS-Language SYstem Evaluation Campaign*, Departement of Computer Science and Information Systems University of Limerick, Limerick, Ireland, 2004. Documents and Linguistic Technology Group.

[24] Günter Neumann and Bogdan Sacaleanu. Experiments on robust nl question interpretantion and multi-layered document annotation for a cross-lanfuage question/answering system. In *Cross Language Evaluation Forum, Results of the CLEF 2004 CROSS-Language SYstem Evaluation Campaign*, LT-Lab, DFKI, Saarbrücken, Germany, 2004.

[25] Lili Aunimo, Reeta Kuuskoski, and Juha Makkonen. Cross-language question answering at the university of helsinki. In *Cross Language Evaluation Forum, Results of the CLEF 2004 CROSS-Language SYstem Evaluation Campaign*, P.O. Box 68, FIN-00014 UNIVESITY OF HELSINKI, 2004. Department of Computer Science, University of Helsinki.

[26] C. de Pablo, J.L. Martínez-Fernández, P. Martínez, J. Villena, A. M. García-Serrano, J. M. Goñi, and J. C. González. miraqa: Initial experiments in question answering. In *Cross Language Evaluation Forum, Results of the CLEF 2004 CROSS-Language SYstem Evaluation Campaign*, 2004.

[27] Sven Hartrump. Question answering using sentecte parsing and semantic network matching. In *Cross Language Evaluation Forum, Results of the CLEF 2004 CROSS-Language SYstem Evaluation Campaign*, University of Hagen(FernUniversität in Hagen) 58084 Hagen, Germany, 2004. Intelligent Information and Communication Systems, University of Hagen.

[28] Luís Costa. First evaluation of esfinge - a question answering system for portuguese. In *Cross Language Evaluation Forum, Results of the CLEF 2004*

CROSS-Language SYstem Evaluation Campaign, Pb 124 Blindern, 0314 Oslo, Norway, 2004. Linguateca at SINTEF ICT.

[29] Enrique Méndex Díaz, Jesús Vilares Ferro, and Davis Cabrero Souto. Cole ad clef 2004: Rapid prototyping of a qa system for spanish. In *Cross Language Evaluation Forum, Results of the CLEF 2004 CROSS-Language SYstem Evaluation Campaign*, Campus de Elviña s/n, 15701 La Coruña(Spain), 2004. Departmento de Computación, Universidade da Coruña.