Doutoramento em Informática

Contextualização e Activação Semântica na Selecção de Resultados em Sistemas de Pergunta-Resposta

José Miguel Gomes Saias

Orientador
Prof. Doutor Paulo Quaresma

2010

Prefácio

Este documento contém a dissertação intitulada *"Contextualização e Activação Semântica na Selecção de Resultados em Sistemas de Pergunta-Resposta"*, entregue em Dezembro de 2009 no âmbito do Programa de Doutoramento em Informática do autor, *José Miguel Gomes Saias*[1], na Universidade de Évora, em Portugal. O autor é Licenciado e Mestre[2] em Engenharia Informática, também pela Universidade de Évora. Actualmente é *Assistente* no Departamento de Informática daquela Instituição.

O orientador deste trabalho é o Professor Doutor Paulo Quaresma[3], *Professor Associado* no Departamento de Informática da Universidade de Évora.

[1] jsaias@di.uevora.pt

[2] A tese de Mestrado foi entregue em 2003 e intitula-se "*Uma Metodologia para a construção automática de Ontologias e a sua aplicação em Sistemas de Recuperação de Informação*" [Sai03].

[3] pq@uevora.pt

Resumo

Para responder a uma pergunta, uma pessoa pode consultar uma enciclopédia ou a *Web*. Na dúvida entre múltiplas soluções, a tendência é escolher a mais apropriada em função da informação disponível.
Esta tese propõe um modelo para Sistema de Pergunta-Resposta de âmbito geral, com técnicas para a avaliação de um conjunto de respostas e tendência para um resultado, que se espera correcto. O comportamento humano é usado como inspiração, nomeadamente no recurso a fontes de informação como complemento ao que se conhece, para encontrar ou reforçar a validade de uma resposta. A *Web*, textos locais, dicionários ou sistemas de informação geográfica constituem fontes de informação com respostas ou elementos que conduzem à solução para a pergunta.
A interpretação de significados possíveis para termos em respostas ou frases tem o auxílio de uma rede semântica, a base de conhecimento inicial. O modelo procura contextualizar respostas candidatas no tempo, no espaço e também numa dimensão semântica. Isto permite a diferenciação dos possíveis resultados, sob várias perspectivas, servindo de base para técnicas de apreciação que despromovem ou reforçam respostas.
A existência de afinidade semântica entre respostas motiva um reforço do respectivo grau de confiança. A avaliação da afinidade semântica tem por base modelos sobre a recuperação de informação na memória humana, vista como uma estrutura associativa onde elementos de informação propagam a sua activação através de relações com outros elementos.
O sistema SENSO foi desenvolvido com a metodologia proposta e a respectiva avaliação reflecte o efeito das técnicas de apreciação de respostas.

Contextualization and Semantic Activation for Answer Selection on Question Answering Systems

ABSTRACT

Looking for an answer, a person may consult an encyclopedia or perform a Web search. To choose among several solutions, one picks the most suitable result, regarding the available information.

In this thesis, a model for open-domain question answering system is proposed, along with answer evaluation techniques for system result improvement. The model is influenced by human behavior. Searching into one or many information resources to complement the existing knowlegde, in order to find or justify an answer, is a reference procedure followed by this model. Web, local texts, dictionaries or geographic information systems are information sources, having answers or elements that lead to the question solution.

Either in answer or text phrases, the interpretation of concept meaning is supported by the starting knowledge base, wich is a semantic network. Candidate answers are contextualized over time, space and a semantic dimension. This organization enables a multiple perspective differentiation over the answer list and supports the answer appreciation process that will promote some of them.

Answers having semantic similarity reinforce each other. This semantic affinity or common meaning evaluation is based on models for information retrieval in human memory where information elements are stored in an associative structure. The spreading activation process performs the propagation of one element's activation level to other semantically related elements.

The SENSO system was developed with the underlying methodology. Its evaluation shows the effect of the proposed answer appreciation techniques.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este capítulo faz uma introdução ao contexto em que se insere o trabalho e enumera os aspectos que estiveram na sua génese. Alguns conceitos mencionados neste capítulo serão descritos com maior detalhe no capítulo 2.

O enquadramento do tema desta dissertação é feito na secção inicial. Na secção 1.2 encontra-se a motivação para o trabalho e os objectivos delineados apresentam-se na secção 1.3. A secção 1.4 descreve a organização dos restantes capítulos deste documento.

1.1 Enquadramento

O ser humano possui qualidades muito próprias que lhe conferem um lugar de destaque entre as espécies do nosso planeta. A sua invulgar capacidade de aprender foi provavelmente o trunfo para a supremacia alcançada entre o reino animal. Através da observação do ambiente envolvente e da comunicação, seja oral entre sucessivas gerações ou escrita e de alcance temporal mais amplo, o Homem aperfeiçoou a aptidão para adquirir informação e transmiti-la aos seus semelhantes.
Como meio estruturante para a comunicação, a linguagem assume um papel preponderante na evolução do ser humano. Juntamente com a percepção, a memória, o raciocínio e a imaginação, a linguagem permite uma aprendizagem para além da que decorre de experiência directa.

A história recente ficou marcada pelo surgimento de uma invenção com grande impacto na vida humana: o computador. Dotado também de uma linguagem própria, distinta da humana por ser simples, precisa e formal, o computador foi inicialmente utilizado em calculo numérico e para execução de tarefas repetitivas em sistemas especializados. Com a miniaturização e evolução em termos de velocidade e capacidade de armazenamento, o computador transformou-se numa ferramenta de aplicação universal, sendo inclusivamente explorado no desempenho de tarefas que simulam o comportamento inteligente dos humanos.
Numa sociedade que cada vez mais produz e requer informação, a Internet representa

uma incontornável plataforma de comunicação. De acordo com o *Internet World Stats*[1], no final de 2007 os utilizadores de Internet em Portugal representam 73.1% da população, enquanto que na União Europeia a 27 a taxa é de 59.9%. Um dos mais populares componentes da Internet é a *World Wide Web*, ou simplesmente *Web*, que permite a qualquer utilizador a publicação de conteúdos e a acessibilidade aos mesmos desde qualquer ponto do mundo.
O formato electrónico diminui custos e facilita a replicação e o armazenamento de documentos, favorecendo a utilização da *Web* em processos de ensino à distância ou a utilização dos seus recursos como apoio ao ensino convencional. Para além de dicionários e enciclopédias, a *Web* possui bastantes conteúdos com informações úteis, incluindo alguns serviços como a tradução automática ou a pesquisa sobre outros conteúdos.
Perante o aumento da quantidade de informação e pela facilidade de acesso à mesma, não são raras as vezes em que a procura de uma informação conduz a diversos assuntos, divergindo do objectivo inicial. As técnicas automáticas de busca e filtragem de informação tornaram-se imprescindíveis e estão hoje ao serviço de qualquer utilizador, para fins profissionais ou particulares. Nem sempre explícitas, essas técnicas podem estar presentes em portais da *Web* ou variadas soluções de software, como clientes de correio electrónico ou gestores de informação pessoal.
A pesquisa por palavra-chave tem sido amplamente estudada na área de Recuperação de Informação. Resumidamente, é um processo pelo qual o utilizador apresenta um conjunto de termos (as palavras-chave) e o sistema devolve uma lista de documentos com base na presença ou relevância dos termos. Finalizada a parte automática, cabe ao utilizador a análise dos documentos obtidos, o que envolve ler e descartar os que não lhe interessam e, eventualmente, detectar a informação procurada no corpo de um daqueles documentos.
Tanto em ambiente profissional como particular, é frequente surgirem questões concretas que requerem respostas concisas. Existem sistemas que aceleram e automatizam a obtenção da informação pretendida, eliminando a intervenção humana na filtragem dos documentos e na leitura de texto. Em vez de retornar o conteúdo integral de um ou mais documentos, um Sistema de Pergunta-Resposta (SPR) aceita uma questão em língua natural, como *"Quem foi Álvaro de Campos?"*, e devolve uma resposta sucinta (neste caso: *"um heterónimo de Fernando Pessoa"*), retirada dos documentos que analisa. A forma de procurar as respostas depende, entre outros factores, da natureza dos dados em que o sistema se baseia. Por um lado, podem ser repositórios com informação estruturada, como bases de dados, XML ou regras lógicas, ou então documentos com informação não estruturada, como é o caso dos textos comuns. Alguns sistemas são especializados num desses procedimentos, enquanto que outros combinam ambos os métodos.

A *Web* é rica em fontes, estruturadas e não estruturadas, de onde se podem extrair respostas plausíveis para as interrogações formuladas. No entanto, a implementação de um sistema automático de pergunta-resposta de uso genérico é um desafio importante, envolvendo técnicas para captar o significado da linguagem utilizada em cada recurso,

[1] http://www.internetworldstats.com

humana ou outra, e relacioná-lo convenientemente com o contexto da questão, processo que nos humanos se desencadeia naturalmente pela inteligência.

1.2 Motivação

Com a expansão dos meios tecnológicos de comunicação, existe cada vez mais informação acessível ao utilizador comum, o que tem levado a um interesse cada vez mais acentuado em sistemas automáticos para auxílio na procura e selecção de informação. Os SPR prestam um serviço completo ao utilizador, uma vez que aceitam interrogações na língua natural do próprio utilizador, escondem a complexidade de procurar e seleccionar a informação e devolvem uma resposta concreta. Teoricamente, este tipo de serviço poupa tempo ao utilizador e melhora a eficiência na busca de informação.

Muitos SPR têm como base um repositório local com informação fiável sobre um domínio específico. Utilizando técnicas estatísticas, linguísticas e regras sobre o conhecimento específico do tema, estes sistemas conseguem bons resultados para o fim a que se destinam e em que foram optimizados. Contudo, a metodologia que seguem não se mostra adequada se o domínio de aplicação for outro.
Se pensarmos na tecnologia como meio facilitador de informação a todas as pessoas, independentemente das diferenças sociais ou culturais, emerge o interesse num SPR independente de qualquer domínio, ainda que os sistemas especializados tenham reconhecida relevância para grupos de interesse comum, áreas técnicas e outros.

A *Web* pode ser vista como um gigantesco repositório de dados em permanente actualização. Tem uma dimensão de tal ordem que se torna impraticável a criação de uma réplica local, para onde se pudessem importar todos os documentos que os milhões de utilizadores espalhados pelo mundo vão criando e modificando. Por outro lado, mesmo que se conseguisse uma cópia global dos conteúdos *Web*, perder-se-ia o dinamismo inerente à distribuição e partilha dos documentos, sendo complicado gerir as actualizações aplicadas aos documentos originais.
A redundância de informação inerente à *Web* constitui uma vantagem potencial, pois diminui a complexidade do processamento linguístico necessário, relativamente à utilização de bases de textos locais e limitadas. Se existirem diversos textos que expressem a mesma informação, cada um com a sua estrutura frásica, a captura automática dessa informação apresentará menos dificuldades. Basta que o sistema consiga interpretar correctamente pelo menos um desses textos.
Com a extensão *Web Semântica*, para além de um enorme volume de documentos com informação não estruturada, a *Web* passa a incluir documentos em que a semântica é expressa numa linguagem máquina, num formato ideal para o processamento automático. A interligação entre ontologias e documentos na *Web Semântica* constitui um modelo partilhado para descrição de conceitos com o objectivo de facilitar a compreensão. A informação semântica sobre conceitos e respectivas relações, formalmente representada em ontologias, pode servir de suporte a processos de inferência automática e auxiliar no

processamento linguístico de textos.

Ao utilizar a *Web* como fonte de informação abrangente e dinâmica, um sistema fica com acesso a recursos heterogéneos, desde os documentos estáticos a serviços de pesquisa ou outros que contribuam para a funcionalidade pretendida. A multiplicidade de respostas potenciais para uma questão, derivadas da *Web* ou de um grande volume de documentos heterogéneos, requer uma capacidade de apreciação que sustente a escolha de uma dessas opções como a resposta de um SPR.

1.3 Objectivos e contribuições previstas

Recorrer a alguém para lhe perguntar algo é uma forma ancestral de obtenção de informação. É possível identificar alguns procedimentos comuns quando uma pessoa é confrontada com uma pergunta, independentemente do tema da mesma. O sujeito interrogado pode saber a resposta de antemão. Pode ser o caso de uma resposta relativa a um facto elementar que o sujeito sabe, ou resultante de uma inferência sobre elementos que o sujeito conhece. Quando o sujeito não sabe a resposta pode ainda recorrer a algum tipo de documentação e eventualmente encontrar a informação procurada. Se não sabe nem encontra documentação útil, a resposta costuma ser *"não sei"*. Outra situação comum é existir dúvida entre duas ou mais respostas. Também aqui o recurso a documentação pode determinar o sentido da resposta.

O objectivo geral deste trabalho é o desenvolvimento de um modelo de Sistema de Pergunta-Resposta inspirado no comportamento humano, nomeadamente no que diz respeito ao recurso a fontes de informação como complemento ao que se conhece, para encontrar ou reforçar a validade de uma resposta. O modelo deve ser independente de qualquer domínio, prevendo técnicas de consulta de uma ou mais fontes para responder a questões sobre assuntos até então desconhecidos. No limite, se não for encontrada uma resposta, o resultado deve ser o que um humano responderia: *"não sei"*. A busca de informação deve ser orientada de acordo com a necessidade de informação suscitada por uma pergunta.
Como objectivo específico, delineou-se a proposta de critérios genéricos de avaliação da adequação de um resultado provisório à pergunta, para diferentes géneros de respostas, que suportem a identificação dos melhores resultados. Nos cenários em que existem N respostas possíveis, pretende-se que o sistema tenha mecanismos para analisar esse conjunto de respostas e convergir para os resultados mais correctos.
Em suma, e tendo em conta os objectivos estabelecidos pretende-se que as contribuições deste trabalho de investigação incluam:

- estudo comparativo do estado da arte na área de SPR
- proposta de um modelo teórico de SPR com inspiração no exemplo humano

- proposta de técnicas para apreciação de respostas, com o objectivo de melhorar o acerto do sistema
- avaliação daquelas técnicas
- desenvolvimento de um SPR

Para se observar, na prática, o efeito do modelo proposto no processo de resposta a perguntas, foi implementado um SPR designado SENSO. Este sistema[2] foi inicialmente pensado para testar apenas a metodologia de selecção de respostas, mas ao longo do tempo ganhou toda a estrutura de um SPR.

1.4 Estrutura da dissertação

Depois deste capítulo introdutório, o capítulo 2 apresenta os principais conceitos relacionados com SPR e descreve o estado da arte neste domínio de investigação, com referência a sistemas pioneiros e a metodologias actuais. No capítulo 3 descreve-se a visão geral do modelo teórico proposto e a metodologia seguida ao longo do processo de resposta a perguntas. Inclui ainda a arquitectura com os principais componentes do modelo e que foram também implementados no sistema SENSO. O quarto capítulo é dedicado à avaliação do modelo. É explicado o processo de avaliação das técnicas de selecção de respostas e indica-se o resultado dessa avaliação. O quinto e último capítulo da dissertação tem as conclusões, com uma apreciação crítica dos resultados e o balanço das contribuições deste trabalho. No final do capítulo 5 apontam-se caminhos possíveis para desenvolvimento futuro, numa perspectiva de continuidade para o trabalho apresentado.

[2] Ao longo deste documento serão frequentemente empregues os termos *modelo* e *sistema*, a propósito do trabalho desenvolvido. O termo *modelo* é uma alusão à proposta teórica deste trabalho e *sistema* é uma referência à sua implementação concreta, o sistema SENSO.

Capítulo 2

Trabalho Relacionado

O segundo capítulo é dedicado ao trabalho relacionado com o modelo proposto no capítulo seguinte, apresentando o estado da arte para esta área. A primeira secção introduz a teoria que serve de base a este trabalho. Na secção 2.2 descrevem-se características relevantes de Sistemas de Pergunta-Resposta e diferentes metodologias aplicáveis neste domínio, a vários níveis. A penúltima secção apresenta um conjunto de sistemas relacionados com o trabalho desta dissertação e que são igualmente pensados para o Português. Em 2.4 faz-se uma síntese do presente capítulo.

2.1 Base teórica

Precedendo a explicação das metodologias seguidas neste tipo de modelos, esta secção descreve os principais conceitos teóricos em que essas metodologias se inserem.

2.1.1 Recuperação de Informação

A possibilidade de transferir capacidades ou valências próprias dos seres humanos a máquinas ou sistemas virtuais tem merecido a atenção do Homem, ao longo dos tempos. Desde a ambição de conceber uma entidade à sua semelhança, em determinados aspectos, à utilidade de implementar um instrumento de trabalho para auxiliar o ser humano, muitas são as motivações apontadas para alicerçar esta área de trabalho.
A informação é, desde sempre, fundamental à vida das pessoas. Perante um problema, a resolução encontrada por um ser humano depende da informação a que o mesmo tem acesso, independentemente de outros eventuais factores circunstanciais, de natureza económica ou outra. Desde o final do século XIX até aos nossos dias, o registo de dados em formato propício para o tratamento automático incrementou bastante. Das experiências com cartões perfurados, em projectos isolados, até à massificação, na actualidade, onde o processamento automático é prática banal, com a criação de qualquer documento em formato digital desde o primeiro instante. Em 1945, Vannevar Bush publicou um artigo onde visionava ou antevia a utilização de uma máquina capaz de armazenar dados de forma organizada, permitindo pesquisas eficientes, como se da memória

humana se tratasse [Bus45]. Alguns anos mais tarde, surgiram os primeiros sistemas automáticos de recuperação de informação.

A Recuperação de Informação (RI) consiste na pesquisa de elementos de informação útil, que satisfaçam requisitos com o interesse do utilizador [MRS09]. É uma área multidisciplinar, envolvendo:

- Linguística - análise do significado expresso pelos dados a pesquisar
- Matemática - definição de modelos estatísticos ou probabilísticos
- Psicologia - teorias sobre o tratamento da informação na memória humana
- Informática - armazenamento dos dados e execução de algoritmos que implementam os modelos

Os sistemas de RI são muito usados, em particular por facilitarem o acesso à informação pretendida quando esta se encontra escondida ou dispersa num grande volume de dados, para o qual a procura manual seria impossível ou extremamente fastidiosa, e certamente demorada.
O elemento de informação relevante que é procurado pode assumir vários tipos. Isto conduz a diferentes especializações deste género de sistemas. Existem sistemas dedicados à Recuperação de Texto e sistemas de Recuperação de Documentos. O primeiro caso é uma forma de RI onde os dados são constituídos por texto livre e a resposta do sistema é um trecho ou passagem existente no texto. No segundo caso, os dados podem corresponder a documentos estruturados ou não estruturados, mas a resposta do sistema é o conjunto de documentos que está de acordo com os critérios de pesquisa.
Outro conceito relacionado com RI é a Extracção de Informação (EI), que vai mais além na busca de conhecimento. A EI assenta na compreensão de textos ou mensagens para depois alcançar os elementos de informação para moldes que representam a necessidade do utilizador.

A Recuperação de Informação tem aplicação nas mais variadas áreas. Um dos exemplos mais populares de sistemas deste género encontra-se nos motores de busca da *Web*. São sistemas de Recuperação de Documentos que encontram uma lista de documentos, ordenada de acordo com a relevância, como resposta à consulta do utilizador. O funcionamento de um sistema de recuperação de documentos envolve quatro fases:

1. encontrar uma representação dos documentos que permita uma busca eficiente
2. interpretar os critérios de procura definidos pelo utilizador
3. pesquisar, comparando a representação da consulta inserida com a representação criada para os documentos
4. devolver uma lista com os documentos que obedecem aos critérios da pesquisa

A primeira fase, onde se indexa o conjunto de documentos, é determinante para o sucesso de todo o processo. As consultas representam a necessidade de informação do utilizador. Podem considerar-se três categorias de consulta:

- palavras-chave: o utilizador apresenta uma expressão com um conjunto de termos relevantes. A pesquisa resulta num conjunto de documentos cujo texto inclui um ou mais daqueles termos. Existem diversos algoritmos de ordenação da lista de documentos, onde um dos critérios é a frequência desses termos [RJ97].

- padrões: consultas onde o critério de selecção consiste na ocorrência de determinado padrão no texto do documento. Um padrão é uma expressão formada por um ou mais caracteres, usualmente relativos a parte de uma palavra, juntamente com símbolos especiais que representam qualquer letra ou qualquer sequência de caracteres. Desta forma, pode exigir-se a ocorrência de palavras com determinado prefixo, sufixo, ou com outras combinações na sua morfologia.

- pesquisa de campos estruturados: em vez de se basear na análise de texto livre, este tipo de consulta é focado em campos estruturados, que possuem valores de forma conhecida (exemplo: data de publicação). Em geral, estas pesquisas têm uma condição do tipo $CAMPO : VALOR$, ou uma disjunção, ou conjunção, entre várias destas condições. Cada condição pode representar um teste de igualdade ou desigualdade.

Alguns sistemas permitem ainda a realização de consultas mistas, com uma parte estruturada e outra parte relativa ao texto livre que forma o conteúdo dos documentos.

Existe muito trabalho em RI, inclusivamente eventos científicos dedicados a este assunto, que visam fomentar o aperfeiçoamento de metodologias. Um desses eventos é a *Text Retrieval Conference*[1] (TREC), uma conferência da comunidade de RI, que tem entre os seus objectivos a avaliação de modelos em larga escala, em particular para metodologias usadas em recuperação de texto. Para sublinhar a importância desta conferência, realça-se o facto de este ser um evento patrocinado pelas instituições americanas: *National Institute of Standards and Technology*[2] (NIST) e *Intelligence Advanced Research Projects Activity*[3] (IARPA).
A primeira edição da TREC teve lugar em 1992. Nos anos seguintes registou um aumento na quantidade de tarefas propostas, tendo também uma subida do número de participantes nessas tarefas, com 22 países representados em 2003. O trabalho apresentado no TREC incide fundamentalmente no Inglês, com algumas tarefas dedicadas a outros idiomas, como o Espanhol ou o Chinês.
Outro grande fórum de trabalho nesta área, que acontece anualmente desde 2000, é o *Cross Language Evaluation Forum*[4] (CLEF). Este evento promove a investigação em RI,

[1] http://trec.nist.gov/
[2] http://www.nist.gov/
[3] http://www.iarpa.gov/
[4] http://www.clef-campaign.org/

com especial ênfase na multiplicidade de idiomas. Existem várias tarefas ou actividades propostas[5], todas relacionadas com RI, mas pensadas para diversos idiomas, não apenas o Inglês. Em algumas tarefas fomenta-se a capacidade multilingue nos sistemas, por exemplo, para aceitar uma pesquisa com termos em Francês e considerar para resposta documentos escritos em Alemão.

No caso dos sistemas que operam sobre a *Web*, os motores de busca, eles lidam com um gigantesco volume de dados, dispersos por documentos de diferentes formatos. A tarefa de indexação é contínua, pois a todo o instante são criados novos documentos e os existentes podem ser actualizados. Os sistemas mais populares, como o Google[6] ou o Yahoo[7], possuem uma infraestrutura tecnologicamente avançada e muito eficiente que visa absorver rapidamente as modificações na *Web*, para minimizar a probabilidade de responder a uma consulta com base em informação desactualizada.
Há, no entanto, necessidades de informação que estes sistemas não conseguem suprir. Quando o utilizador precisa de respostas concretas, a recuperação de documentos não é suficiente. É nestes casos que se usam os Sistemas de Pergunta-Resposta, o tema tratado na secção seguinte.

2.1.2 Sistemas de Pergunta-Resposta

A Inteligência Artificial (IA) é uma ciência dedicada ao estudo de modelos e algoritmos para o desempenho automático de funções que os seres humanos realizam com recurso à sua inteligência. Aplicando o poder computacional sobre o processamento de símbolos, procuram-se métodos genéricos para automatizar actividades perceptivas, cognitivas e manipulativas [Per88]. Engloba a criação de mecanismos, em particular programas de computador, cujo aspecto inteligente é a capacidade de atingir objectivos concretos [McC98].
O Processamento de Linguagem Natural (PLN) é um ramo da IA, e que envolve também a área de Linguística, relacionado com o tratamento automático de linguagens naturais ao Homem, na forma escrita ou falada. Esta compreensão automática da linguagem natural passa por mecanismos de análise do discurso, de resolução de ambiguidade e de conversão, de e para, linguagens formais que são próprias dos algoritmos.
Algumas actividades de PLN são: sumarização automática, tradução automática entre idiomas, reconhecimento de entidades mencionadas, resolução de anáforas, conversão de discurso falado para texto ou vice-versa, e também a tarefa de responder automaticamente a uma pergunta.

Um Sistema de Pergunta-Resposta (SPR) tem por objectivo responder automaticamente a questões formuladas em língua natural aos humanos, fornecendo um resultado

[5] http://www.clef-campaign.org/2009/2009_agenda.html
[6] www.google.com
[7] www.yahoo.com

conciso, resultante da consulta de um texto ou de outra fonte de informação. A aplicabilidade deste tipo de sistemas é muito abrangente, desde um sistema de ajuda *online*[8] com informação sobre peixes de aquário até aos sofisticados sistemas de apoio ao negócio, de informação clínica para a área da saúde e farmácia, ou mesmo a análise de dados estratégicos militares [May03].
Os SPR podem responder às perguntas com base em fontes de informação estruturada, como bases de dados sobre um domínio específico, ou fontes de informação não estruturada, por exemplo baseada em textos. Muitos dos sistemas actuais recorrem a fontes de ambos os géneros, diversificando a proveniência das respostas [Mon03].

A ideia de usar máquinas para responder a perguntas escritas surgiu há muito. Em 1950, Turing descreve uma tarefa designada *Imitation Game*, onde dois intervenientes ocultos respondem por escrito às perguntas de uma pessoa. Turing propõe a substituição de um desses intervenientes humanos por uma máquina, questionando o impacto dessa troca no juízo efectuado sobre a identidade dos participantes ocultos [Tur50]. Alguns anos mais tarde, surgiram os primeiros SPR dedicados a simular uma conversação com o utilizador [Sim65], respondendo a qualquer pergunta com base na sua estrutura sintáctica. Alguns exemplos populares para este tipo de sistema são o CONVERSATION MACHINE [GBG59] e ELIZA [Wei66]. Este último, simulava uma conversa com o psicólogo. Funcionava com base na detecção termos chave no discurso do interlocutor humano, a partir dos quais, e através de regras definidas por especialistas no domínio, mantinha a conversa activa.
Outros SPR recorrem a bases de dados para responder a interrogações sobre factos de um domínio específico. Permitem um acesso simples ao conteúdo da base de dados, funcionando como interface entre o utilizador e o repositório estruturado com os factos, que consultam usando uma linguagem de interrogação própria, como o SQL[9]. São exemplos deste género os sistemas BASEBALL [GWCL61] e LUNAR [Woo73, Woo77]. O primeiro fornece respostas sobre jogos de *baseball*, acerca de encontros disputados, respectivos locais, resultados e estatísticas. As equipas e o elemento pretendido eram identificados no texto da questão, seguindo-se uma consulta à base de dados com todos os resultados da época desportiva. O sistema LUNAR fornece informações sobre rochas lunares. Este sistema consulta uma base de dados sobre amostras recolhidas na Lua durante a missão *Apollo 11*. A descrição de ambos os sistemas aponta para bons resultados na resposta a perguntas sobre a respectiva área de especialização. Não podem, contudo, encontrar a resposta a matérias que não são abrangidas pela base de dados de suporte.
Um terceiro género de SPR, entre os sistemas pioneiros, testava a compreensão automática de um texto. Em 1975, Schank apresentou o sistema MARGIE, capaz de analisar e interpretar um texto, para depois responder a perguntas sobre o seu conteúdo [RSGR75]. O sistema baseava-se em teorias sobre a organização da memória humana e na inferência sobre uma representação semântica criada para o texto analisado.

[8] Acessível através da Internet.
[9] SQL: *Structured Query Language*, é uma linguagem de definição, modificação e consulta de dados usada em Sistemas de Gestão de Bases de Dados.

Os primeiros SPR são anteriores à ARPANET[10]. Com o crescimento das redes de computadores e o acesso generalizado à Internet, os SPR são influenciados de duas maneiras. Por um lado, os sistemas passaram a comunicar com máquinas remotas, aumentando o número e o tipo de fontes de informação. Depois, a facilidade de acesso aos SPR através das redes veio aumentar o leque de potenciais utilizadores.

O processo de responder automaticamente a uma pergunta é bastante complexo. Encontrar documentos relacionados com a pergunta, ou com o assunto a que se refere a pergunta, é apenas uma parte do processo. Apesar de existirem diferentes metodologias, a maioria dos SPR realiza um conjunto de tarefas gerais que é comum:

1. **análise da pergunta** - interpretação da pergunta para determinar o assunto a que se refere, ou seja o alvo ou foco da questão. Importa apurar também o que pede a pergunta sobre o seu foco, por exemplo uma característica de valor numérico, uma data ou a definição. Esta interpretação assenta fundamentalmente na análise morfo-sintáctica do texto da pergunta, em particular a verificação dos pronomes e entidades mencionadas, e vai direccionar o processo de procura de possíveis respostas.

2. **recuperação de documentos** - encontrar documentos que possam conter uma resposta à pergunta. A partir do foco da questão e outros elementos da pergunta, gera-se uma expressão com palavras-chave para consulta numa ferramenta que executa recuperação de documentos. Muitos sistemas têm como única fonte de respostas o resultado desta ferramenta de recuperação de documentos, que costuma ser uma lista ordenada por ordem decrescente de relevância face à consulta. Se houver problemas na formação da consulta ou na execução da mesma na ferramenta de recuperação de documentos, então o resultado do sistema fica desde logo comprometido.

3. **análise do documento** - para cada documento identificado na fase anterior realiza-se uma verificação do seu conteúdo, em busca de entidades mencionadas no texto da pergunta, de ocorrências do foco da pergunta e de expressões aceitáveis, do ponto de vista morfológico e semântico, como possíveis respostas. Alguns sistemas escolhem passagens do texto que consideram relevantes, por incluírem expressões naquelas condições, e passam-nas à fase seguinte para um nível mais apurado de processamento.

4. **extracção de respostas** - De cada expressão assinalada extrai-se uma resposta candidata, constituída pelo valor da expressão, uma parte dessa expressão, ou um valor calculado em função da expressão. Alguns sistemas associam um valor

[10]ARPANET: Rede de computadores que esteve na base da actual Internet, desenvolvida com o apoio da *Advanced Research Projects Agency* (ARPA). Em 1969 interligava 4 universidades americanas. A partir de 1970 expandiu-se para fora da América, ligando também Londres e a Noruega. A partir de 1990 passou a designar-se simplesmente Internet. Um ano depois surgiu a World Wide Web.

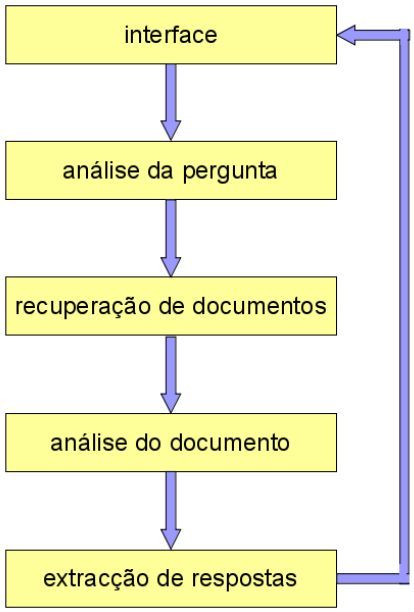

Figura 2.1: Arquitectura genérica de um Sistema de Pergunta-Resposta

numérico a cada resposta candidata, com o peso ou grau de confiança que lhe é atribuído. Quando o sistema devolve apenas uma resposta, escolhe-se a resposta candidata que no final do processo apresentar o maior grau de confiança.

A implementação de cada tarefa pode seguir variadas abordagens. Para responder a uma pergunta, executam-se estas tarefas pela ordem apresentada, como ilustra a figura 2.1, com a arquitectura geral resumida de um SPR. A interacção entre utilizador e SPR realiza-se através de um interface, cuja especificação e complexidade varia de sistema para sistema.

Na última década tem havido muito trabalho em SPR. As principais conferências relacionadas com RI incluem geralmente uma actividade, ou mesmo uma oficina paralela, especialmente dedicada a esta área de investigação. É o que acontece nas conferências mencionadas antes. O evento TREC tinha uma ramo de Pergunta-Resposta, o TREC-QA[11], cuja última edição decorreu em 2007. Associado ao CLEF existe também a actividade QA@CLEF[12], que engloba a tarefa principal de Pergunta-Resposta de domínio aberto, sobre colecções de textos, e ainda uma série de actividades mais específicas, incluindo a validação de respostas. Na última edição da tarefa principal do QA@CLEF estiveram em avaliação sistemas mono e multilingue, envolvendo um total de 10 línguas

[11]TREC-QA: http://trec.nist.gov/data/qamain.html
[12]QA@CLEF: http://clef-qa.itc.it/

faladas no continente europeu. Outro grande evento, o NTCIR[13], promove também um conjunto de actividades relacionadas com RI e SPR, para linguagens asiáticas.

2.2 Características e metodologias de SPR

O tratamento das perguntas e a procura de respostas seguem, em termos abstractos, uma sequência de fases comuns a muitos sistemas. Mas a implementação concreta depende, entre outros factores, dos objectivos de cada sistema. A análise às características dos SPR pode realizar-se sob vários aspectos:

1. o tipo de respostas devolvidas
2. a área de conhecimento em que consegue dar respostas
3. as fontes de informação que usa para responder
4. o processo utilizado para responder

Existem sistemas especializados em respostas factuais [CA05], que nesta área de trabalho costumam ser designadas simplesmente por factóides. Estas são respostas curtas, cujo valor pode, por exemplo, corresponder a uma cor, um lugar, o nome de uma pessoa ou um valor numérico. Outros sistemas permitem respostas mais longas [Niu07], as descrições. Este é o resultado de questões que pedem alguma caracterização, por exemplo as definições. A identificação precisa do início e fim de uma resposta deste tipo no meio de um texto pode ser mais complexa que a identificação dos factóides. A outro nível, há ainda as respostas do tipo lista. Independentemente do comprimento da resposta, o que a caracteriza é ser formada por uma colecção de itens pretendidos para o resultado de uma pergunta como: *"Quais as freguesias do concelho de Évora?"*.

Relativamente à área de conhecimento em que os SPR operam, existem sistemas de domínio fechado ou restrito e sistemas de âmbito geral ou domínio aberto. Os primeiros são especializados em determinado tópico, por exemplo as estatísticas de um campeonato nacional de futebol ou os procedimentos de manutenção de um automóvel, ou uma interface em língua natural para qualquer sistema pericial. Os sistemas BASEBALL e LUNAR, referidos na secção anterior, são exemplos de SPR de domínio fechado. Um SPR diz-se de domínio aberto se está preparado para responder a perguntas sobre qualquer assunto. Para tal, é necessário que a metodologia do sistema permita a procura de respostas em função do tópico associado à pergunta. Este tipo de sistemas ganhou fôlego com o surgimento e a rápida expansão da *Web*, pois o acesso a documentos sobre os mais variados temas passou a ser rápido e simples. Em contrapartida, nestes casos o volume de dados a consultar é muito grande e pode incluir documentos contraditórios ou imprecisos.

A resposta a uma pergunta tem origem numa fonte de informação que pode ser um texto, uma base de dados, um dicionário ou outro recurso que o sistema possa analisar.

[13]NTCIR: http://research.nii.ac.jp/ntcir/ntcir-ws8/ws-en.html

Algumas fontes são estruturadas, o que significa que podem ser consultadas automaticamente com uma linguagem formal, como o SQL. Noutros casos a informação não é estruturada, como acontece com os textos. A informação encontra-se expressa em língua natural e a sua compreensão já não se realiza de um modo directo.
As fontes de informação podem ser locais ao sistema ou externas. As fontes locais podem ser repositórios de conhecimento num formato próprio para o mecanismo de inferência do sistema, podem ainda ser bases de dados ou colecções de textos. As fontes externas são independentes do sistema, como por exemplo a *Web*, uma enciclopédia ou dicionário *online*. Nestas fontes, o armazenamento, a actualização e o controlo dos dados são alheios ao SPR, que pode apenas realizar consultas, geralmente através da rede e seguindo um protocolo específico.
Os primeiros SPR utilizavam apenas fontes locais. Mais tarde, os sistemas de domínio fechado passaram a usar também serviços remotos sobre a respectiva área de especialização. Os sistemas de âmbito geral actuais tendem a usar a *Web* como fonte de respostas, por terem aí documentos de suporte para os mais inusitados temas.

Em traços gerais, a metodologia de resposta dos SPR modernos pode enquadrar-se nas seguintes categorias:

- **análise superficial de textos** - a extracção de respostas é feita com base na ocorrência de determinados padrões no texto. São aplicadas regras com expressões regulares de casos típicos de resposta para determinado género de perguntas [GS04]. As respostas captadas por este processo podem passar depois por uma análise estatística que contribua para a identificação da resposta mais plausível.

- **análise linguística** - aplica técnicas de PLN que visam captar a semântica de frases em textos consultados pelo sistema e também da própria pergunta. Deste modo, a extracção de respostas faz-se com maior precisão, avaliando a compatibilidade entre a interpretação da pergunta e a compreensão (possível) dos textos de suporte. Esta metodologia é mais morosa que a primeira, por aplicar um conjunto de operações necessárias à análise textual, como o reconhecimento de entidades mencionadas, resolução de anáforas ou a normalização de variações morfo-sintácticas.

- **raciocínio** - para além de usarem a metodologia anterior, há sistemas que encontram algumas respostas por um processo mais complexo, que envolve algum raciocínio. Isto acontece quando uma resposta não é resultante de uma frase, mas antes da soma de um conjunto de evidências, que são detectadas em diferentes documentos, possivelmente associados a fontes de informação diferentes e de vários tipos. É necessário um formalismo para representar o conhecimento que o sistema tem, juntamente com o que pode descobrir nas fontes que consulta. A interpretação da pergunta, nesse formalismo, através de um mecanismo de inferência pode conduzir a resultados que a metodologia anterior não alcança.

Num plano mais concreto, há um vasto conjunto de técnicas empregues em SPR, no tratamento da pergunta, em colecções de textos locais ou sobre elementos de informação

obtidos no processo de resposta. Algumas destas técnicas são enumeradas em seguida.
A classificação das perguntas é a identificação da categoria de cada pergunta, logo na fase de análise inicial, o que dará pistas sobre o tipo de resposta a encontrar. Já tinha sido referida a necessidade de determinar o foco da questão. A categoria vem complementar essa informação, representando o género de interrogação efectuada sobre aquele foco. Para cada categoria de pergunta podem usar-se estratégias distintas de resolução.
Em 1977, Wendy Lehnert elaborou um modelo teórico chamado QUALM, muito citado na bibliografia de SPR. É um modelo para resposta a perguntas no contexto da compreensão de histórias, realizado no âmbito de um doutoramento em filosofia. A compreensão da pergunta e a procura pela resposta são dependentes do contexto e da noção pragmática do que constitui uma resposta aceitável. Na metodologia QUALM, para o Inglês, existiam 13 categorias de questões [Leh77] para testar a compreensão de textos, entre as quais:

- antecedente causal: *Why did John kill the dragon?*
- consequência: *What happens if John leave?*
- especificação de características: *How old is John?*

No entanto, este não é um SPR completo. A implementação existente funciona associada a outros sistemas, requerendo a existência de um grafo conceptual tanto para a pergunta como para a informação de base para a resposta: o texto com a história.

O sistema MURAX aplica uma metodologia baseada na análise linguística e enquadramento ou plausibilidade das respostas candidatas relativamente à categoria da pergunta [Kup93]. Este sistema tem como fonte de informação uma enciclopédia *online*. A enciclopédia é consultada em dois momentos no processamento de uma pergunta. Primeiro procuram-se documentos relevantes, para a categoria de pergunta e respectivo foco, dos quais se extrai um primeiro grupo de respostas candidatas. A segunda pesquisa na enciclopédia é uma apreciação de cada resposta candidata, consumada pela pesquisa de hipóteses que misturam texto da questão com a resposta.
Outro sistema onde a classificação das perguntas representa uma parte importante do processo é o WEBCLOPEDIA [HGH+00]. As categorias a atribuir a uma questão foram organizadas numa taxonomia onde os tipos mais gerais estão perto da raiz e os mais concretos são as "folhas". Com base na observação de milhares de perguntas e respectivas respostas, a cada categoria foram associadas indicações de moldes ou padrões textuais (válidos para o Inglês), que permitem captar uma resposta candidata, e um peso em função da precisão estimada para esse molde [HHR02]. Um dos exemplos dessa taxonomia, designada pelos autores *QA Typology*, é a categoria R-POPULATION. Para uma pergunta como *"What is the population of Venezuela?"*, essa categoria sugere a aplicação de moldes como:

```
0.60  <NAME>' s  <C-QUANTITY> population
0.37  <NAME>' s  <C-QUANTITY> population
0.37  of <NAME>' s  <C-QUANTITY> people
```

Em cada linha, a primeira coluna indica o peso daquele padrão. A etiqueta <NAME> representa o foco da questão, *Venezuela*. A última etiqueta representa a resposta a extrair, aplicando uma ferramenta de expressões regulares sobre o texto.
Alguns modelos estendem a noção de categoria semântica também às fontes de informação. Yun Niu desenvolveu uma metodologia para responder a perguntas descritivas, não factóides, para a área clínica [Niu07]. Definem-se as categorias semânticas para aquele domínio em particular, caracterizadas por um conjunto de propriedades para as quais se registará uma polaridade (positiva, neutra ou negativa). Em seguida, o repositório local com textos médicos é analisado para se apurar quais as categorias abrangidas por cada documento, bem como a polaridade nas respectivas propriedades. A pergunta é analisada por um processo semelhante. As respostas são obtidas com base nas relações existentes nas categorias semânticas da pergunta e a restante base de conhecimento.

Ao utilizar uma ferramenta de recuperação de documentos para encontrar a informação necessária, existe o risco de não encontrar documentos, ou obter alguns que se revelam insuficientes para que a fase de extracção encontre resultados. À semelhança de outros sistemas que utilizam recuperação de documentos, os SPR usam a técnica de expansão da expressão de pesquisa para motores de busca, sejam locais ou para a *Web*. A pesquisa costuma ser baseada em palavras-chave, com a possibilidade de especificar partes opcionais e mandatórias, e de formar disjunções ou conjunções. Esta expansão consiste em enriquecer a expressão de pesquisa, acrescentando-lhe alternativas semânticas para alguns termos. As alternativas resultam de dicionários, ontologias e redes semânticas. Podem ser alcunhas, sinónimos, hipónimos ou outras especializações dos conceitos na expressão original. Para esta operação de expansão semântica dos conceitos a pesquisar, o sistema WEBCLOPEDIA, a par de muitos outros SPR para língua inglesa, consulta o recurso WordNet [Mil95]. Mas o aperfeiçoamento da expressão de consulta nem sempre corresponde à inserção de termos. Por vezes retiram-se termos da expressão inicial, quando estes são considerados pouco informativos e demasiado restritivos [Bil04]. O trabalho [ALG01] descreve um estudo sobre as transformações a efectuar sobre a expressão de pesquisa inicial, para os SPR obterem melhores resultados dos motores de busca.

A maioria dos sistemas tenta minimizar a área de texto em que tem de aplicar as técnicas mais morosas. Assim, após a recuperação de documentos, é comum aplicar-se a técnica de recuperação de trechos[14], para seleccionar partes específicas de cada documento. A implementação da recuperação de trechos pode ser mais ou menos elaborada. Uma solução simples consiste em dividir um texto em parágrafos e escolher aqueles em que ocorrem determinadas palavras-chave. Em versões mais sofisticadas, formam-se várias janelas de texto a partir do documento original, eventualmente sobrepostas. Essas janelas de texto são depois analisadas por uma ferramenta de recuperação de informação, obedecendo a um modelo vectorial ou probabilístico [RWJ+95, Hu06], que

[14]Recuperação de trechos ou passagens no texto é uma operação referida na bibliografia em Inglês com o nome *passage retrieval*.

dará uma ordenação dos trechos. Para terminar, o sistema selecciona os trechos mais relevantes, que estarão no topo da lista.

O reconhecimento de entidades mencionadas (REM) é uma importante tarefa de PLN que os SPR aplicam ao texto dos documentos consultados e também sobre a pergunta. Por um lado, esta técnica ajuda a interpretar correctamente a pergunta ou frases com nomes compridos ou designações de entidades formadas por várias palavras. Para além disso, a aplicação prévia de REM sobre uma colecção de textos local permite orientar a procura de documentos relevantes no processamento de uma questão, fornecendo apontadores para os locais onde surgem entidades de determinada categoria. As classes de entidades encontradas por este processo incluem: pessoa, organização, local, data, hora, percentagem ou quantia monetária. Ao trabalhar com REM sobre uma colecção de textos local, o cruzamento de dados sobre entidades específicas, de todos os documentos que mencionam essa entidade, pode conduzir à descoberta de características sobre cada entidade, o que é de grande utilidade para um SPR [Sar03]. A catalogação de nomes de pessoas, por exemplo, pode ajudar a descobrir a sua data de nascimento, a cidade onde vive ou a empresa em que trabalha.

Nos SPR de âmbito geral, conseguir um bom desempenho nos resultados de vários tipos de perguntas obriga não somente à consulta de diversas fontes como também ao uso de múltiplas estratégias de resolução [KLL+03]. O sistema PRISM, por exemplo, combina técnicas reconhecidamente úteis, usadas por sistemas mais antigos, com a utilização da *Web* [Wu03]. Quando aplicadas sobre uma fonte de informação fiável, muitas vezes um conjunto local de documentos, as tradicionais técnicas de análise linguística conseguem uma apreciável precisão na identificação das respostas. Mas a complexidade do processo deixa escapar algumas respostas, que podem ocorrer num só local em todo o repositório de informação. Os sistemas que procuram respostas na *Web* dispõem de mais documentos onde procurar. Como há uma grande probabilidade da mesma informação se encontrar escrita de várias formas e em diversos documentos, estes sistemas têm vantagem na captação de respostas candidatas. Em contrapartida, a menor fiabilidade dos documentos usados, relativamente a uma fonte local controlada, leva a uma menor precisão nas respostas. O sistema PRISM procura tirar partido do melhor da cada abordagem, com a abrangência da *Web* a favorecer a extracção e a análise linguística sobre fontes credíveis a assegurar a precisão possível, no sentido de melhorar a qualidade geral dos resultados.
O procedimento descrito acima é uma forma de projecção de respostas candidatas. Esta técnica consiste em recolher respostas candidatas numa fonte onde a extracção é mais fácil, não necessariamente a *Web*, seguida de uma análise da validade de cada resultado, para a pergunta em causa, numa outra fonte. Ao consultar a segunda fonte no tratamento de uma pergunta, existe informação sobre a resposta em apreciação, o que facilita a detecção de eventuais elementos que comprovem essa resposta.
Um sistema mais recente que também usa projecção de respostas é o QUALIM[15] [Kai08].

[15] http://demos.inf.ed.ac.uk:8080/qualim/

As respostas extraídas em diversas fontes, incluindo documentos *Web* encontrados via Google, são 'projectadas' sobre uma segunda fonte, a Wikipédia. O QuALiM apresenta no resultado apenas as respostas candidatas que encontraram eco na Wikipédia, inseridas num parágrafo obtido daquela fonte, juntamente com um apontador para o documento.

A *Web* é um recurso vasto e dinâmico, com características próprias que condicionam os SPR, mas que por outro lado também dão lugar à utilização de novas abordagens. Algumas técnicas de análise linguística intensiva, válidas para colecções isoladas com documentos, não são escaláveis para o volume de dados da *Web*. Seria impraticável manter um índice sobre os conteúdos a nível global. A redundância existente na *Web* propicia a utilização de técnicas estatísticas para a extracção de respostas [MNPT02a]. Para chegar aos documentos redundantes acerca daquilo que o sistema pretende, usam-se os motores de busca existentes na *Web*, como o Google ou o Altavista.
O recurso à *Web* visa melhorar o desempenho do sistema, mas existem dificuldades que, potencialmente, contrariam os benefícios teóricos dessa fonte de respostas. Num tutorial sobre técnicas usadas em SPR da *Association of Computational Linguistics*, em 2003, Jimmy Lin enumerava algumas dessas dificuldades [LK03].:

- o comportamento dos motores de pesquisa pode variar no tempo
- o volume de dados captados que não são úteis é muito superior aos documentos com respostas correctas
- dificuldades no processamento linguístico para resolver anáforas em textos obtidos da *Web*, sem prejudicar demasiado o tempo de resposta do sistema
- a existência de documentos antigos na *Web*, cujas respostas serão desactualizadas
- o valor mais popular nem sempre é o mais correcto
- a verificação de expressões regulares para extracção de respostas de determinados padrões textuais é por vezes falaciosa, quando os textos são relativos a mitos, ficção ou ironias

A presença de erros nos textos e a heterogeneidade nas fontes são aspectos que podem ocasionar respostas erradas. Mas espera-se que esses valores tenham um peso residual quando comparado com uma resposta correcta, detectada num documento e corroborada em muitos outros.

O primeiro SPR a ser disponibilizado na *Web* foi desenvolvido por Boris Katz, no *Massachusetts Institute of Technology*, e designa-se START (*SynTactic Analysis using Reversible Transformations*). A respectiva interface[16], que presta o serviço a utilizadores de qualquer parte do mundo, está activa desde Dezembro de 1993 [Kat97]. Antes de possuir

[16]Endereço do sistema START: http://start.csail.mit.edu/

uma interface *Web* era já um SPR com vários anos de desenvolvimento que explorava padrões lexicais na língua inglesa, [Kat88, KL88]. Inicialmente, o sistema era composto por dois módulos complementares, originalmente designados *understanding module* e *generating module*. O primeiro faz análise de textos em Inglês e gera uma representação para a informação, que acumula num repositório de conhecimento. O segundo módulo constrói frases em língua natural a partir de elementos de informação no repositório. Usados em conjunto, permitem interpretar a questão do utilizador e devolver-lhe uma resposta igualmente em língua natural. A representação de cada frase consiste numa estrutura ternária (originalmente chamada *T-expression*) com os elementos principais identificados na estrutura sintáctica dessa frase. Em geral, a estrutura guarda tuplos *(Sujeito,Relação,Objecto)*, onde a relação corresponde ao verbo da acção, na maior parte dos casos, ou a alguma preposição em situações especiais. Cada estrutura pode ter, no sujeito ou no objecto, outra estrutura ternária. É desta forma que são representadas frases com várias orações, subordinante e subordinadas. Estes tuplos são ainda associados a expressões em língua natural, designadas anotações, que facilitam a reconstrução das frases no momento de responder com dados destas estruturas, para além de ajudarem na procura de recursos relacionados com o mesmo tópico. As respostas são encontradas pela inferência realizada a partir dos tuplos. O processo é controlado com um conjunto de regras para a interpretação semântica de verbos e respectivos argumentos, para o Inglês. Estas regras, *S-rules*, especificam por exemplo quando é que um tuplo equivale a outro, estando um na voz passiva e outro na voz activa. A notoriedade conquistada pelo sistema atraiu diversos parceiros e levou a sucessivos aumentos do repositório de conhecimento, nomeadamente a partir do *CIA World Factbook*[17] e da análise de variados recursos na *Web*.

OMNIBASE é uma base de dados virtual que aglomera dados de diversas fontes heterogéneas, estruturadas e não estruturadas, constituindo um repositório de conhecimento uniforme [KFY+02]. Funciona como uma interface para um leque de fontes de informação, simplificando o acesso aos dados. A informação está organizada de acordo com um modelo *Objecto-Propriedade-Valor*, onde cada tuplo representa um facto registado. Esta estrutura simples permite a execução de consultas de um modo normalizado, independente dos detalhes na fonte original dos dados. Esta plataforma serve actualmente de suporte ao sistema START.

Outro sistema popular, que dispõe de uma interface *Web*[18], é o ASKJEEVES. Tem por base um grande repositório de anotações sobre o conteúdo de documentos *Web* e consultas por palavra-chave. Apesar de rápido, este sistema responde com uma ou mais frases, o que constitui um resultado mais longo que o esperado num SPR.

O sistema MULDER executa múltiplas consultas no motor de busca Google [KEW01]. Essas consultas são determinadas após a análise sintáctica e subsequente classificação da pergunta. A partir do verbo da pergunta e da categoria que lhe é atribuída, com

[17] Repositório de informação sobre a história, população, economia, geografia e outros aspectos relativos a muitos países. https://www.cia.gov/library/publications/the-world-factbook/

[18] AskJeeves: www.ask.com

recurso à WordNet[Mil95], o sistema realiza a expansão da interrogação formulada para a consulta ao motor de pesquisa. Sobre os documentos *Web* resultantes, aplica técnicas de PLN e um conjunto de heurísticas para a extracção de respostas. Cada resposta candidata é pontuada em função da sua posição no texto, relativamente a determinadas palavras-chave. Após a graduação individual, as respostas candidatas são agrupadas em função das palavras que possuem em comum. Cada grupo é então avaliado, sendo-lhe atribuída uma pontuação igual à soma da pontuação de cada resposta que contém. A resposta candidata de maior pontuação em cada grupo é designada de resposta representativa. O sistema termina o processamento da pergunta mostrando ao utilizador a resposta representativa de cada grupo, numa lista ordenada.
Os autores do sistema MULDER usaram uma métrica de avaliação sobre o esforço necessário para o utilizador até conseguir a resposta à sua pergunta. Para isso, contabiliza-se o número de palavras que o utilizador terá de ler, desde o início do conteúdo apresentado como resultado do sistema até à expressão que constitui a resposta certa. Na experiência realizada, sobre a taxa de esforço para o utilizador, os autores anunciam que o sistema requer menos 6,6 vezes esforço de leitura que o necessário através do Google.

Outro exemplo de sistema que privilegia a redundância dos dados, deixando a análise linguística para segundo plano, é AskMSR [BDB02]. A metodologia, em termos abstractos, é semelhante aos sistemas deste género: reformulação da pergunta num conjunto de interrogações para motores de busca na *Web*; extracção de respostas com base na ocorrência de determinado padrão textual.
As consultas realizadas sobre a *Web* contém expressões que resultam na reformulação da pergunta num sentido declarativo ou assertivo. As regras para gerar estas expressões são definidas manualmente. Um exemplo indicado na descrição deste sistema refere que a pergunta *"When was the paper clip invented?"* tem como reformulação a expressão *"The paper clip was invented"*, que poderá ser prefixo de uma resposta candidata.
As expressões geradas para a interrogação têm diversos níveis de precisão. Isso é assinalado, proporcionalmente, num valor numérico que representa o peso dessa variante de reformulação da pergunta. No nível mais baixo de precisão, a interrogação corresponde a uma conjunção dos termos da pergunta que não são *stop words*[19]. Ao procurar respostas candidatas sobre o resultado da recuperação de documentos, o sistema tem em conta o peso da interrogação usada para obter cada documento. Esse peso constitui um parâmetro na determinação da pontuação de cada resultado.
Neste sistema, considera-se que não responder é mais aceitável que responder incorrectamente. Para tentar reduzir as respostas erradas, o sistema usa uma técnica de predição de erro que filtra as respostas candidatas com reduzida pontuação, ou cuja estrutura não seja adequada ao tipo de resposta previsto.

Para responder a perguntas sobre factos recentes publicados em notícias, o sistema AnswerBus[20] indexa conteúdos *Web* de `www.cnn.com`. O sistema de indexação usado

[19] *Stop words* são palavras muito frequentes e de pouco valor informativo. Exemplos: *que*, *de*, *e*.
[20] http://www.answerbus.com

é parte importante do sistema, incluindo por exemplo a noção de tempo de cada artigo daquele arquivo histórico noticioso [Zhe03]. A resolução de anáforas é uma das técnicas de PLN usadas. As questões são classificadas em 19 categorias, incluindo *Definição* e *Exemplo*. O sistema admite perguntas escritas em língua natural para seis idiomas, incluindo o Português. Ao detectar uma questão numa língua diferente do Inglês, o sistema traduz automaticamente a pergunta para depois prosseguir com o processo de resolução, suportado por fontes com conteúdos em Inglês.

A existência de recursos em idiomas diferentes, por exemplo na *Web*, dá lugar a situações em que uma pergunta, redigida na língua A, pode ter respostas em documentos escritos em línguas diferentes, B e C, não sendo encontrado nenhum resultado nas fontes da língua A. Para melhorar os resultados em situações como a descrita, alguns sistemas usam técnicas multilingue. Outro cenário que justifica capacidades multilingue é permitir a um utilizador apresentar a pergunta na sua própria língua, por exemplo em Português, a um SPR cujas fontes se encontram em Alemão. Aqui, o sistema poderia traduzir automaticamente a pergunta, por exemplo. A dissertação de Sergio Escámez [Esc08] descreve uma arquitectura multilingue, com participações na tarefa bilingue Inglês-Espanhol do QA@CLEF, e enumera dificuldades associadas à tradução automática, que pode alterar o sentido da pergunta.

As respostas mais surpreendentes são aquelas que não resultam de uma correspondência óbvia entre uma frase nas fontes do sistema e a pergunta, mas antes da inferência a partir de elementos de informação que isoladamente não levariam a nenhum resultado. Para conseguir tais respostas, é necessária uma base conhecimento onde os dados de cada fonte são expressos numa linguagem formal, adequada a um mecanismo de inferência. Um sistema nestes moldes foi usado por Quaresma & Rodrigues no QA@CLEF de 2005 [QR05]. Aplicando técnicas de PLN, o sistema gera uma estrutura lógica representativa do discurso. Uma estrutura semelhante é construída para a pergunta, para ser depois interpretada por um mecanismo lógico de inferência, baseado na linguagem declarativa PROLOG.

Após a fase de extracção de respostas candidatas, é comum realizar-se uma validação para eliminar respostas com menor grau de confiança. Esse grau de confiança é atribuído na fase de extracção ou calculado posteriormente, através de consultas às fontes do sistema, que combinam a pergunta e a resposta candidata.
Chin-Yew Lin desenvolveu uma técnica de reordenação de resultados para as perguntas do tipo *Definição*, com o propósito de aumentar o grau de confiança das melhores respostas [Lin02], composta por duas variantes. A primeira variante revê a pontuação de cada resultado em função do valor estimado de concordância com uma definição de referência, obtida no recurso WordNet. A segunda variante desta técnica consiste na pesquisa do termo a definir, usando três motores de busca. Os resultados de cada consulta *Web* são analisados para extracção de uma janela de texto, em redor do termo a definir. A comparação das palavras na resposta candidata com as palavras de cada janela de texto vai

determinar a bonificação para essa resposta. Ambas as variantes foram testadas com o sistema WEBCLOPEDIA para um conjunto de definições do TREC-QA, tendo-se registado uma melhoria nos resultados.
Em 2005, Dalmas & Webber propuseram um modelo de fusão de informação para ordenação de respostas candidatas a que chamaram QAAM [DW05]. Para este modelo, é formado um grafo cujos nós representam palavras da pergunta e respostas candidatas. Os elos que ligam os nós podem denotar *equivalência* ou *inclusão*, sendo estabelecidos em função das relações sinónimo, hiperónimo e merónimo, consultadas no WordNet para os termos desses nós. A ordenação das respostas faz-se em função de propriedades da partição do grafo em que se inserem os respectivos nós. Este modelo de grafo com os dois tipos de relação entre os nós foi aplicado a respostas com locais e definições.
A metodologia aplicada melhora os resultados de sistemas que à partida têm um bom conjunto de respostas candidatas. Mas nas situações onde os resultados de uma pergunta são maioritariamente errados, ou onde os resultados correctos são desagregados e há uma grupo de resultados errados mas relacionados, então esta metodologia não é adequada [DW07].
Em [KSN07] a selecção das respostas obedece a um modelo probabilístico de avaliação de cada resultado. Para cada resposta candidata é calculada a probabilidade da mesma ser válida separadamente em cada fonte de informação usada, nomeadamente o WordNet e a Wikipédia. Entre cada par de respostas candidatas, calcula-se ainda a probabilidade das mesmas serem semanticamente semelhantes. No final, estima-se uma probabilidade da pergunta ser correcta, em função dos valores anteriores.

2.3 Sistemas de PR para o Português

O trabalho em SPR para a Língua Portuguesa aumentou nos últimos anos, talvez devido à visibilidade que alguns fóruns de avaliação como o QA@CLEF trouxeram para esta área de investigação.
O PERGUNTE foi um dos primeiros sistemas a usar a *Web* em Português como fonte de respostas [RB05]. Desenvolvido por um grupo da Universidade Federal de Pernambuco, no Brasil, o sistema recebe perguntas em língua natural, classifica-as numa categoria, realiza pesquisas nos motores de busca e extrai respostas através da análise superficial dos textos. Com o resultado da análise morfo-sintáctica ao texto da pergunta, o sistema classifica cada questão numa das classes: *Localização, Data, Quantidade, Razão, Porque_Famoso, Modo, Definição, Abreviação, Abreviação_Expansão, Função e Nome.*

Uma equipa da Universidade de Lisboa desenvolveu o XISQUÊ, um SPR de domínio aberto sobre a *Web* e para o Português [BRSS08]. Este sistema assenta na metodologia desenvolvida num trabalho anterior chamado QUEXTING, para a resposta a perguntas factuais [Rod07]. Pensado para responder em tempo real, através da sua interface *Web*[21], o XISQUÊ começa por analisar a questão em busca do verbo principal e palavras

[21] http://xisque.di.fc.ul.pt/

relevantes. Essas palavras-chave são usadas para uma consulta em motores de busca, sendo os documentos resultantes descarregados para processamento local. A extracção de respostas no XisQuê começa pela identificação das frases que poderão conter um resultado, tendo em conta o esperado para a pergunta. Em seguida, as frases seleccionadas são analisadas em busca de respostas tipificadas como *"respostas curtas"*. Se o sistema não consegue identificar esse tipo de resposta, então a própria frase é usada como *"resposta longa"*. As respostas fornecidas pelo sistema nem sempre são tão concisas quanto o desejável para um SPR. Por exemplo a questão *"Quem é Nelson Évora?"* é respondida com frases onde ocorre o nome *Nelson Évora*. O requisito de responder em tempo real é apontado como um dos aspectos chave do sistema. O tempo médio para responder a 60 perguntas do jogo *Trivial Pursuit* foi 21 segundos, sendo 64% desse período gasto na interacção com o exterior através da rede.

Os sistemas que se seguem têm um historial de participação em actividades de avaliação de metodologias de resposta. Independentemente de usarem ou não a *Web* ou outros recursos auxiliares, todos respondem a perguntas com base em fontes de texto locais.
O sistema raposa foi usado pela Faculdade de Engenharia da Universidade do Porto na edição de 2008 do QA@CLEF [STO08]. A sua metodologia começa por identificar o tipo de pergunta e determinar as palavras-chave que formam a base da expressão de pesquisa para a recuperação de documentos. Nesta primeira fase, são utilizados o analisador lexical JSpell [AP95] e o siemês [Sar06] para o reconhecimento de entidades mencionadas. A expansão da expressão de pesquisa é a fase do processo mais enfatizada pelos autores, com o objectivo de melhorar a recuperação de documentos e contribuir para aumentar a abrangência dos resultados devolvidos. Para perguntas cuja resposta esperada é um factóide, o elemento central é uma acção ou evento. A selecção de frases relevantes para encontrar a resposta é dependente de verbos que possam ter a ver com essa acção ou evento. Assim, a expressão de pesquisa terá o verbo original e a respectiva expansão, formada por verbos equivalentes ou relacionados. A extracção de respostas candidatas faz-se por dois processos:

- regras de análise superficial de texto que procuram padrões associados ao contexto da pergunta
- identificação de expressões cujo tipo se enquadre no género de resposta procurada e que ocorrem com maior frequência nas passagens seleccionadas

Para além dos componentes da arquitectura convencional, inclui um módulo de fusão das respostas candidatas. Apesar de não estar implementado na versão de 2008, este módulo visa o agrupamento de respostas lexicograficamente distintas mas com significados idênticos.

O próximo sistema deve o seu nome a uma história da mitologia Antiga acerca de um demónio que estrangulava pessoas que não conseguissem resolver os seus enigmas. Chama-se esfinge e é desenvolvido por uma equipa da Linguateca. É um SPR de

âmbito geral com cinco participações no QA@CLEF e pode ser consultado através de uma interface *Web*[22]. Com um processo de selecção de documentos e extracção de respostas candidatas baseado na procura de padrões textuais, a principal fonte de respostas deste sistema é a *Web* [CC08]. Usa o analisador morfo-sintáctico PALAVRAS[Bic00] e o reconhecimento de entidades mencionadas é feito com o SIEMÊS. Na versão de 2008, o sistema é composto por um conjunto de módulos sequenciais [Cos08]:

- resolução de anáforas: substituição do elemento anafórico detectado na estrutura sintáctica da pergunta por candidatos existentes no contexto (perguntas do mesmo tópico ou a resposta da primeira pergunta do grupo). À questão inicial juntam-se estas questões alternativas, nas quais o elemento anafórico está resolvido.

- reformulação da questão: especificação dos critérios de procura, que seguem duas abordagens para a pesquisa de textos com potenciais respostas. Primeiro, o sistema colecciona trechos efectuando procuras superficiais no texto. O processo seguinte é a procura com base na estrutura sintáctica da pergunta. O analisador PALAVRAS é aplicado a cada frase e são selecionados os casos onde o verbo principal no texto é uma conjugação do verbo principal na pergunta. Se o processo de extracção de respostas não encontra resultados nos textos existentes, então realiza-se ainda uma selecção de trechos (ST) com base na estrutura sintáctica, mas sem a exigência do verbo principal da pergunta.

- extracção de respostas nos trechos encontrados: aplicando sequencialmente os procedimentos

  1. padrões textuais
  2. reconhecedor de entidades mencionadas
  3. selecção de N-gramas (sequências de palavras)

- filtragem de resultados: são descartadas expressões que fazem parte da pergunta ou que correspondem a palavras muito frequentes

- associação da resposta a um texto de suporte: procurar um trecho de texto local diferente da origem da resposta candidata (em particular se provém da *Web*) que confirme o resultado. Esta é uma tarefa necessária no QA@CLEF.

- apresentação da resposta

Com duas participações no QA@CLEF, o sistema QA@L2F é um SPR de âmbito geral, criado em 2007, e é focado no processamento linguístico [CMG+08]. A sua metodologia tem três fases principais:

[22] http://www.linguateca.pt/Esfinge

- pré-processamento das colecções locais de textos: os documentos locais ao sistema, que são a origem das respostas, são previamente processados. A aplicação da análise linguística a todos os textos motiva esta fase inicial. O procedimento inclui detecção de entidades mencionadas e a aplicação de um módulo de resolução de anáforas. Este tratamento inicial das fontes de informação vai popular uma base de dados que será consultada nas fases posteriores, nomeadamente para a selecção de trechos onde procurar respostas.

- interpretação da pergunta: a questão é analisada e transformada num estrutura usada para a selecção de trechos de texto relevantes. Esta selecção inclui a realização de consultas em SQL. A estrutura inclui o tipo da pergunta, o foco da mesma, entidades mencionadas e outros elementos como o verbo principal.

- extracção de respostas

    1. pela recolha de entidades mencionadas que respeitem o tipo previsto para a resposta e que se encontrem próximas ao foco da questão ou de palavras-chave na estrutura descritiva da pergunta
    2. pela procura superficial de padrões no texto para os documentos locais da Wikipédia

O sistema IdSay é o mais recente entre os descritos, tendo marcado presença apenas na edição de 2008 da tarefa principal do QA@CLEF em Português [CdMR08]. Entre os componentes da arquitectura do sistema, que é a convencional, os autores realçam a componente de recuperação de informação, que procura aliar a eficácia à independência da língua usada nos documentos a consultar. O IdSay permite a escolha, para cada pergunta, da colecção local de documentos a considerar como fonte de respostas. A análise inicial classifica a questão numa categoria (*Factóide*, *Definição* ou *Lista*) e gera uma expressão de pesquisa com entidades mencionadas e palavras relevantes para a recuperação de documentos. Os textos encontrados passam por um componente de selecção dos trechos que poderão conter uma resposta. Cada trecho tem um comprimento até 60 palavras, que pode ser excedido para completar frases, no início ou na parte final do trecho. As respostas candidatas que passam pela validação são apresentadas como resultado para a pergunta.
O componente de indexação dos textos locais foi implementado em camadas, com os procedimentos que dependem da língua dos documentos separados dos procedimentos gerais, tendo em vista uma futura aplicação multilingue. A extracção de respostas consiste no reconhecimento de entidades mencionadas nos trechos e na aplicação de regras de análise superficial de padrões textuais. Quando existem duas respostas muito próximas, no mesmo trecho de texto, elas são fundidas, formando uma só resposta que inclui ambas e ainda o texto que as separa.

Por último, o sistema que tem alcançado os melhores resultados gerais no QA@CLEF

em Português: PRIBERAM. Este é um SPR desenvolvido por uma empresa[23] especializada em conteúdos e ferramentas linguísticas, nomeadamente dicionários, correctores ortográficos e motores de busca. É um sistema de domínio aberto com uma metodologia de intensa análise linguística. Participando no QA@CLEF há já cinco anos, e em diversas actividades[24], este sistema apresenta uma arquitectura formada por cinco fases: fase de indexação, análise da pergunta, recuperação de documentos, recuperação de frases e extracção de respostas [ACF+08].
O SPR PRIBERAM dispõe de um elaborado mecanismo de indexação de texto. É realizada uma meticulosa análise ao nível da frase (e não apenas do documento), que lhe associa desde logo um conjunto de categorias de pergunta para as quais poderão ter relevância, e ainda uma lista com os domínios conceptuais (relativos a uma ontologia) que estão associados a essa frase. A recuperação de trechos de texto tem em consideração o tipo da pergunta e domínios conceptuais relevantes para a pergunta, seleccionando as frases cujo pré-processamento tenha assinalado elementos compatíveis.
Quando uma questão é classificada numa categoria, os padrões textuais de selecção de respostas, próprios dessa categoria, são aplicados sobre as frases que o processo de recuperação selecciona para a mesma categoria. A análise à questão é realizada com o recurso FLiP[25] e um reconhecedor de entidades mencionadas. O foco da questão e outros elementos nucleares são identificados pela estrutura sintáctica. O uso de uma ontologia de conceitos a que ficam associadas as frases contribuiu para o aumento de precisão na recuperação de trechos, o que se reflecte também nas respostas.

No último capítulo, a tabela 5.1 faz um resumo de algumas características destes sistemas, lado a lado com o sistema desenvolvido no âmbito deste trabalho.

2.4 Resumo

A Recuperação de Informação (RI) ajuda-nos a encontrar elementos de informação dispersos por colecções de dados e cujo acesso seria difícil de concretizar por pesquisa sequencial. A Recuperação de Documentos e a Recuperação de Texto são formas de RI que permitem encontrar, respectivamente, documentos e excertos de texto, mediante um critério de procura. Um Sistema de Pergunta-Resposta aceita perguntas na língua natural aos humanos, por exemplo em Língua Portuguesa, e devolve uma ou mais respostas concisas, obtidas automaticamente por consulta em fontes de informação, ou alcançadas através de um processo de raciocínio.
A arquitectura genérica de um SPR inclui os componentes: análise da pergunta, recuperação de documentos das fontes de informação, análise do conteúdo do documento e extracção de respostas. Os SPR podem operar num domínio específico, utilizando recursos com informação precisa sobre o respectivo tema. Outros sistemas respondem

[23] http://www.priberam.pt
[24] Este sistema participou também no QA@CLEF em Espanhol.
[25] FLiP: Ferramentas da Língua Portuguesa - http://www.flip.pt

a perguntas de vários tipos e sobre qualquer tema. Estes são os sistemas de domínio aberto. A metodologia para encontrar respostas pode ser mais superficial (às vezes complementada com algum tipo de análise estatística, de validação ou de projecção de respostas) ou suportada por uma análise linguística, que é mais morosa mas permite extrair respostas com um maior grau de informação.
Os sistemas existentes para o Português seguem metodologias com algumas partes em comum mas também com algumas diferenças nas técnicas empregues. Alguns recorrem fundamentalmente à *Web*, outros não o fazem de todo, diferenças que se observam também nos SPR mais populares para a Língua Inglesa.

Capítulo 3

O Modelo Proposto

O modelo teórico proposto nesta dissertação é apresentado neste capítulo. A primeira secção introduz uma sequência de aspectos chave que servem de base para o modelo. As técnicas propostas para atingir os objectivos previstos para o modelo são descridas na secção 3.2. Em 3.3 explica-se o funcionamento do sistema de pergunta-resposta associado a este trabalho e as circunstâncias em que tira partido deste modelo. A penúltima secção, 3.4, recapitula os componentes do sistema desenvolvido e o modo como interagem. O capítulo encerra com uma síntese do conteúdo exposto, em 3.5.

3.1 Visão geral

A intenção fundamental do trabalho aqui descrito é dotar um sistema de pergunta-resposta de âmbito geral com a faculdade de escolher de modo acertado e fundamentado entre as diversas opções de resposta, algo que nos humanos se considera um comportamento inteligente. Reconhecendo que, na prática, se afigura como uma utopia a concretização de tal capacidade num sistema completamente automático e para qualquer domínio do conhecimento, usa-se essa visão como linha de orientação para o modelo.
Para melhorar a capacidade de escolha de um sistema, o modelo procura reproduzir alguns fenómenos associados à mente humana, em particular a nível da associação de factos ou informações em memória. Usando uma analogia metafórica, pretende-se que um sistema de pergunta-resposta assuma comportamentos que teria um humano, eventualmente estrangeiro e com conhecimento rudimentar da Língua de base do sistema. Genericamente, a atitude inicial é tentar compreender a pergunta, recorrendo a formas de tradução para a língua que domina. Havendo dificuldades no entendimento de um termo, é necessário perceber que conceitos é que lhe estão semanticamente associados e em que contextos é que esse termo é empregue. Estas duas medidas podem determinar o juízo de um humano não falante da língua base, no sentido de optar ou não por uma resposta que inclua estes termos mal compreendidos.
Para além de consultar dicionários, efectuar pesquisas na *Web* e em textos disponíveis, interessa imitar também o modo de associação de ideias no caso humano, o modo de

Figura 3.1: Objectos para exercício de agrupamento

interpretação da ligação semântica entre conceitos e contextualização de experiências, usando tudo isso na selecção das respostas do sistema.

3.1.1 A Linguagem e a Memória

O Homem assimila facilmente a Linguagem, resultado do desenvolvimento social e cultural, e serve-se dela para interagir e para codificar experiências vividas ou ideias não concretizadas. O acto de escutar ou ler uma pergunta e em seguida responder desencadeia-se nas pessoas com grande naturalidade. Não obstante, a compreensão da pergunta, escrita ou falada, constitui um processo neurolinguístico muito complexo. Para um humano, ler uma palavra ou escutar os fonemas que a compõem leva à activação de determinados circuitos em redes semânticas associativas que existem nosso cérebro. Estes circuitos estabelecem ligação entre palavras e definições, conceitos associados e ainda vivências ou acontecimentos relacionados, contribuindo para uma correcta interpretação das frases [DMP07] [Lam99].

No início da década de 1930, o psicólogo soviético Alexander Luria levou a cabo um importante estudo sobre a influência sócio-cultural no desenvolvimento cognitivo. As experiências realizadas com habitantes da Ásia Central mostram que o processo cognitivo varia em função do modo com que os diversos grupos sociais vivem e experimentam diversas realidades [Lur76]. Uma das tarefas realizadas pelos indivíduos em estudo era a classificação de objectos, sendo-lhes mostrado um quadro com quatro elementos de onde teriam de excluir um, aquele que não se enquadraria no grupo dos outros três, e justificar essa opção. Um dos quadros continha um serrote (de metal), um machado, um martelo e madeira (blocos de madeira), como ilustrado na figura 3.1.

Ao avaliar os resultados verificou-se que havia grupos de pessoas que excluíam:

- a madeira, por considerarem os restantes como um grupo de ferramentas
- o serrote, por não incluir madeira na sua constituição
- o machado, associando os restantes a actividades de carpintaria

- o martelo, encarando os restantes como madeira e utensílios para o respectivo corte

Note-se que todas as respostas são plausíveis. Deste estudo, evidenciam-se duas formas básicas de que a mente humana faz uso durante o processamento dos conceitos, nomeadamente numa tarefa de classificação de objectos. Por um lado, há o tratamento por categoria, que tem a ver com a semântica do conceito em causa e envolve generalização a partir de elementos concretos. Por outro lado, há ainda o tratamento pelo contexto em que o objecto se insere, o que depende das experiências quotidianas de cada sujeito. Os dois últimos tipos de resposta são disso exemplo. Muitas das justificações eram alusivas a situações vividas por aquelas pessoas e que envolviam madeira, serra e uma das outras ferramentas.
Observou-se também que parte significativa destas respostas por agrupamento situacional eram relativas a pessoas sem escolaridade ou com poucos anos de formação, sendo a resposta por categoria dada por indivíduos com mais anos de estudos, mais habituados a exercícios que envolvem abstracção.

Um ser humano possui estruturas cognitivas que associam conceitos a contextos concretos ou ao respectivo tipo ou significado geral. Em psicologia e neurociências distingue-se Memória Episódica de Memória Semântica, duas memórias do tipo explícito que fazem parte da nossa Memória de Longo Prazo. Enquanto a primeira tem a ver com recordações sobre eventos e vivências onde vários conceitos se relacionam em determinado contexto, a segunda tem um carácter mais geral, associando conceitos aos seus significados [Ing92]. Uma das primeiras teorias sobre a retenção de informação na Memória de Longo Prazo foi o modelo de Memória Semântica Hierárquica [CQ69], de *Collins* e *Quillian*. Neste modelo, o conhecimento organiza-se em unidades de informação ou conceitos interligados de forma hierárquica, formando uma rede semântica. Alguns anos mais tarde, *John Anderson* apresentou a teoria *Adaptive Control of Thought* [And83], com a noção de grau de activação associado a cada unidade cognitiva ou nó da rede semântica. A recuperação de informação da memória de longo prazo realiza-se por um processo de propagação de activação (*spreading activation process*) iniciado nos nós de informação presentes na chamada memória de trabalho. Estes possuem ligações a outros nós existentes na memória de longo prazo. A força da ligação é medida pelo grau de activação: quanto maior o grau, maior será a probabilidade de conseguir rapidamente o resultado do processo de pesquisa. O grau de activação pode aumentar ou diminuir no tempo devido a vários factores, nomeadamente a frequência de pesquisas ou activações da ligação.

Estabelecendo um paralelo com o campo das neurociências e com os tipos de memória que permitem ao ser humano identificar, contextualizar e relacionar conceitos, este modelo recorre também a repositórios de informação de natureza distinta. Na perspectiva semântica, utiliza-se a base de conhecimento inicial do sistema, que constitui uma rede de conceitos e relações estabelecidas entre eles, incluindo hiperónimo, merónimo, sinónimo e outras, como se descreve adiante.
Na vertente episódica, a associação contextual entre conceitos e entidades é detectada no conteúdo de fontes de informação não estruturada. Este género de associações, que

as vivências de uma pessoa vão acumulando na memória, corresponde neste modelo ao que o sistema capta em narrações de factos, num conjunto de textos em língua natural, locais ao sistema, ou provenientes da *Web*.

3.1.2 Contextos semântico, espacial e temporal

O ambiente em que se insere cada indivíduo influencia o modo como interpreta uma pergunta. Existem questões sobre um tópico cuja área temática se encontra bem explícita. A resposta a tais questões deve encontrar-se nessa mesma área semântica. Outras questões podem ser relativas a um conceito que faz sentido em distintos contextos, não sendo claro a qual deles se referem.
Um pessoa tem a noção de espaço e tempo. Algumas respostas para a mesma questão podem variar em função da localização, do tempo ou de ambos. É importante que o sistema seja também sensível a estas dimensões no tratamento de uma pergunta.

Contexto semântico

A ambiguidade é uma dificuldade associada à língua natural. Na pergunta:

O que é um ensaio ?

o conceito a definir é *ensaio*, que pode ter os significados:

- primeira prova ou tentativa de alguma coisa
- esboço literário ou científico
- jogada de râguebi em que se coloca a bola no chão, atrás da baliza adversária

Como a pergunta não especifica um domínio em particular, todas as respostas são válidas, cada uma no seu contexto. É possível orientar o sistema para determinado contexto. No exemplo, podemos acrescentar "*em râguebi*" ou indicar *desporto* como interesse do utilizador. Esta indicação ou pista funciona como uma preferência. O sistema identifica e dá maior peso às respostas compatíveis com o contexto pretendido.

Contexto espacial

Durante uma conversa entre duas pessoas, o ambiente envolvente contextualiza o discurso e dá significado a expressões de localização relativa ou indirecta. A mesma pergunta pode ter diferentes respostas, em função da localização de quem responde. Como exemplos podemos considerar as questões:

- *Onde estás?*
- *A que distância fica Lisboa?*

- *Quem é o Presidente?*

Para responder de forma adequada, o sistema precisa de uma localização base, como se fosse uma pessoa a habitar numa cidade em concreto. Esta localização serve de referência para encontrar o espaço a que a pergunta se refere.
Nas duas últimas questões no exemplo, poder-se-ia ainda considerar que as respostas tanto podem ser relativas à localização de quem pergunta como de quem responde. Neste modelo, estas questões que dependem de um contexto espacial omisso interpretam-se em função de quem responde, com base na localização do sistema. A cidade de Évora foi escolhida como local em que o sistema se encontra.
Outras questões têm associada uma expressão que restringe as respostas a determinada região ou zona geográfica, como acontece em:

- *O que é o IPM em Portugal?*
- *Que cidade francesa foi residência dos Papas?*
- *Em que zona de Espanha nasce o rio Douro?*

Aqui a localização do sistema não é relevante. As expressões *em Portugal*, *cidade francesa* e *zona de Espanha* dão pistas sobre o contexto geográfico a ter em conta para encontrar a resposta.

Contexto temporal

Outra dimensão importante neste modelo é o tempo. A resposta a uma pergunta pode variar em função do tempo, inclusivamente quando a referência geográfica é a mesma. São exemplo disso as perguntas:

- *Que horas são?*
- *Quem é o Presidente da República?*
- *Qual é a idade de Cavaco Silva?*

O sistema usa a data e hora actuais como momento presente. Se a pergunta não é relativa a uma data específica, então o sistema dá prioridade à resposta mais próxima da actualidade (o que acontece na segunda questão do exemplo). A última questão apresentada ilustra o caso em que é necessário considerar o tempo decorrido. Para encontrar a idade pedida, o sistema calcula os anos passados desde a data de nascimento de Cavaco Silva, se essa data for conhecida por algum meio. Se o sistema detectar uma afirmação sobre a idade, como *"Cavaco Silva tem X anos"*, poderá então somar X ao número de anos decorridos desde a data daquela afirmação, gerando assim mais uma resposta.
A percepção temporal no modelo compreende diferentes momentos:

- momento actual - representa o presente e permite ao sistema calcular o tempo que passou desde outra data (um nascimento, uma inauguração ou data de outro evento).

- tempo da acção - identifica a data de um acontecimento ou acção. Esta informação pode encontrar-se numa frase como a seguinte:

  O Presidente da República tomou posse em 9 de Março de 2006.

  A frase é explícita quanto à data da acção *tomar posse*. Ao analisar o texto, o sistema capta esta data e usa-a como resposta ou como referência para uma resposta.

- tempo da narração dos factos - representa o momento em que a acção é relatada. Nos textos noticiosos corresponde à data em que a informação é publicada. É especialmente útil para resolver expressões temporais relativas (como *ontem*, *dia 10 do mês passado*, *na próxima quinta-feira*) e assim deduzir o tempo da acção.

- tempo da resposta - estabelece um critério para a selecção ou para o cálculo da resposta a uma pergunta, com base no tempo da acção ou no tempo da narração dos factos. Na pergunta:

  Quantos anos tinha Cavaco Silva em 2004?

  o ano 2004 é o tempo da resposta. Para obter a idade naquele contexto temporal, o sistema pode:

  - Encontrar uma afirmação sobre a idade de *Cavaco Silva* em 2004. Tal afirmação pode constar de um documento desse mesmo ano ou num documento posterior (ou mesmo anterior) mas em que o tempo da acção é 2004.
  - Encontrar uma referência à idade de *Cavaco Silva* num ano anterior a 2004 e somar o número de anos até ao tempo da resposta.
  - Encontrar uma referência à data de nascimento de *Cavaco Silva* e calcular o número de anos até ao tempo da resposta.

Percepção do contexto

O modelo em estudo tem um âmbito geral, não sendo dedicado especialmente a um contexto semântico. Tendo a data actual como referência dinâmica para o presente, o sistema está associado a uma localização em Évora. Dessa localização concreta, é possível deduzir o contexto geográfico mais abrangente, Portugal, e com maior abstracção, a Europa e depois o mundo.
Os dados que suportam uma resposta são extraídos de fontes de informação, estruturada ou não, mas que possuem, muitas vezes, meta-informação com o tema do texto ou a data de publicação. Quando detectados pelo sistema, esses elementos são considerados no momento em que um texto é usado para suportar uma resposta.

Tanto para os textos de suporte como para as perguntas, o sucesso na detecção do contexto semântico, espacial ou temporal é influenciado pela acuidade da análise linguística do sistema.

3.1.3 Base de conhecimento inicial

Consideremos um conjunto de palavras $\{p1, p2, ...pn\}$. Suponhamos que não sabemos o significado de nenhuma das palavras do conjunto. Compreender uma frase, isolada, constituída por estas palavras seria uma tarefa muito complicada, mesmo para uma pessoa. Seria necessário recorrer a ferramentas como dicionários. Para um sistema automático também é necessário algum tipo de recurso que forneça pistas ou significados para os termos a processar.
A base de conhecimento inicial no modelo é uma rede semântica que inclui conceitos de diversos domínios e relações estabelecidas entre os mesmos. As principais relações semânticas são enumeradas em seguida.

1. **hiperónimo** - expressa uma relação de especialização entre categorias ou conceitos. Exemplos: *mamífero* é hiperónimo de *felídeo*; *árvore* é hiperónimo de *oliveira*.

2. **merónimo** - ser parte constituinte de algo. Exemplos: *dedo* é merónimo de *mão*; *ilha* é merónimo de *arquipélago*.

3. **sinónimo** - denota equivalência no significado dos termos que associa. Exemplo de sinónimos: *hesitante*, *indeciso* e *reticente*.

4. **antónimo** - relação usada para assinalar significados contrários. Exemplos: *quente* é antónimo de *frio*; *avançar* é antónimo de *recuar*.

5. **relacionadoCom** - associação entre conceitos que podem estar relacionados num mesmo campo de ideias. Exemplo: *bombeiro* está relacionado com *incêndio*.

6. **instânciaDe** - relaciona uma categoria semântica com uma sua concretização ou entidade que a materializa. Exemplo: *Évora* é uma instância de *cidade*.

7. **localizadoEm** - associa dois locais ou conceitos entre os quais se pode estabelecer um posicionamento relativo. Exemplos: *Évora* situa-se em *Portugal*; *missa* tem como local a *igreja*.

8. **agente de determinado verbo** - usada para assinalar instâncias ou categorias tipicamente praticantes de determinadas acções. Exemplos: *polícia* é agente de verbo *disparar*; *pássaro* é agente de verbo *voar*.

Estes conceitos e respectivas relações formam uma rede semântica cuja informação é estruturada e fácil de consultar. Estão incluídos alguns factos acerca de lugares, entidades e eventos. Esta rede semântica servirá de suporte a processos de inferência e análise semântica.

A rede semântica funciona como uma base de conhecimento de senso comum. O seu conteúdo é uma combinação de elementos importados de estruturas existentes e alguma edição manual. A hierarquia de conceitos de topo deriva da *Ontologia Senso*, inspirada em trabalho anterior sobre ontologias [SQ05] [ACQS05] e que beneficiou da importação de alguns factos de senso comum acerca de assuntos quotidianos, do repositório do sistema *ConceptNet* [LS04], de acordo com o descrito em [SQ06]. Mais recentemente foram introduzidas algumas relações provenientes da ontologia PAPEL [OSGS08].

3.1.4 Fontes de respostas

A resposta a uma pergunta pode ser um valor numérico, uma data, um nome, uma expressão curta ou uma longa descrição. O resultado a encontrar pode estar na base de conhecimento inicial do sistema ou pode ser externo ao motor de inferência, proveniente de um recurso local ou remoto. A procura de possíveis respostas e elementos que sustentam essas respostas faz-se na *Web*, em textos noticiosos, em ontologias, em documentos locais, em dicionários e sistemas de informação geográfica.

Web

Consultar a *Web* para encontrar informação tornou-se usual no quotidiano das pessoas. Um estudo recente [SAD08] sugere que, face a uma pergunta, o recurso por parte do utilizador a serviços de pesquisa na *Web*, de modo interactivo em busca documentos que conduzam a uma resposta, pode alcançar resultados aproximados aos obtidos por sistemas completamente automáticos. Estes sistemas, em geral, têm vantagem na velocidade com que realizam um grande número de pesquisas sobre os mais diversos temas, perdendo para os humanos na precisão das respostas.
Pela abrangência dos seus conteúdos e por estar em permanente actualização, a *Web* foi escolhida como uma das principais fontes de respostas para o modelo. Trabalhos anteriores evidenciaram o aspecto positivo da redundância inerente à *Web*. A existência de várias representações para a mesma informação facilita a obtenção de resultados com técnicas simples de pesquisa de padrões [CCL01] [KLL+03] [LK03].
Para cada tipo de pergunta existe um procedimento próprio para selecção de documentos *Web*, através de motores de busca comuns para os quais é preparado um padrão de procura. Para além de uma pesquisa global sobre a *Web*, há casos em que se realizam também pesquisas mais circunscritas, sobre enciclopédias e arquivos de jornais na Internet. As enciclopédias são consultadas pelo sistema por constituírem uma fonte credível, com grande variedade de factos unanimemente aceites. A utilização dos textos de carácter noticioso tem dois aspectos relevantes para o modelo. Por um lado, as notícias contém factos sobre assuntos mediáticos, que são potenciais alvos de perguntas. Depois, a existência de notícias recentes, designadamente as informações que vão sendo publicadas hora a hora, conferem algum dinamismo ao sistema, em particular a capacidade de alterar o sentido de uma resposta em tempo real.

Ontologias

A *Web Semântica* veio estender a *Web* tradicional com um conjunto de normas para a representação de informação de modo propício para o processamento automático. Para além dos textos, imagens, formas e cores, destinados primordialmente à leitura por humanos, os documentos são enriquecidos com anotações ou meta-informação. Algumas normas da *Web Semântica* prevêem a utilização de ontologias, que são referidas em documentos, tornando-os semanticamente mais ricos e contribuindo para que os conceitos lá existentes sejam melhor entendidos.
Nos sistemas de domínio fechado é já usual a procura de respostas numa ontologia, como forma estruturada e especializada de representação de conceitos e relações daquele domínio alvo. Assim acontece, por exemplo, na metodologia de Ferrández et al [FEIBFEVG08].
Um sistema de pergunta-resposta de âmbito geral não impõe restrições ao domínio ou ao léxico empregue nas questões que trata. O texto presente nessas questões e nos documentos analisados engloba um enorme conjunto de conceitos, referidos por termos que o sistema deve interpretar correctamente. Como meio estruturante de informação, as ontologias são também consideradas neste modelo. Para além de auxiliarem na análise de relações entre conceitos, por exemplo a relação de especialização, uma ontologia pode conter directamente a resposta a uma questão.

Documentos locais

O modelo inclui uma vasta colecção local de textos sobre diversos assuntos. Grande parte dos documentos provem de uma cópia da enciclopédia Wikipédia[1]. Existem textos com notícias da década de 90, de arquivos de jornais de Portugal e do Brasil. E há ainda documentos diversos, alguns dos quais inseridos directamente por utilizadores.
Ao prever a existência desta colecção de textos, cujo acesso se faz directamente pelo sistema, o modelo permite o seguinte:

1. **utilização de um motor de pesquisa personalizado**
   A pesquisa via *Web* condiciona o sistema às opções do motor de busca usado. Na pesquisa a nível local o sistema recorre a um motor de busca especialmente adaptado aos documentos. O processo de indexação e pesquisa pode beneficiar de recursos linguísticos personalizados, nomeadamente na verificação de sinónimos ou obtenção do termo raiz de uma palavra. Se a colecção local for multilingue, uma pesquisa pode recorrer a ferramentas optimizadas para cada idioma.

2. **processamento linguístico dos textos pode fazer-se de antemão**
   A análise morfológica e sintáctica a textos, escolhidos como potencialmente relevantes para uma questão, é morosa e pode comprometer o tempo de resposta

[1] http://pt.wikipedia.org/

do sistema, que se pretende rápido. A análise linguística aos textos locais pode efectuar-se previamente, sem prejuízo para o normal funcionamento sistema. No caso dos documentos *Web*, a análise textual mais detalhada incide apenas sobre determinadas frases ou parágrafos escolhidos pelo sistema, também para não prejudicar demasiado o tempo da resposta.

A utilização destes documentos permite ainda a obtenção de resultados mesmo quando, pontualmente, o sistema não consegue aceder à *Web* nas melhores condições. Cada documento tem associada uma informação com a data de edição ou publicação. Esta indicação permite organizar cronologicamente os textos relativos a um mesmo assunto e será usada para contextualizar no tempo as respostas oriundas desses documentos.

Dicionários

O modelo recorre a dicionários para verificação de sinónimos, o que é útil em diversas fases do processamento de uma pergunta. Mas um dicionário pode também fornecer uma resposta efectiva, designadamente para questões do tipo definição, se a expressão a descrever corresponder a uma entrada do dicionário.
O acesso a dicionários faz-se por duas formas. A versão mais directa consiste na consulta de uma tabela local. Esta tabela pode ser adaptada, em particular para inserção de novas entradas. Para além disso, o sistema acede também a dicionários de livre acesso na *Web*.

Sistemas de informação geográfica

Saber onde ficam monumentos, rios, cidades e muitas outras entidades com existência espacial requer a consulta de repositórios de informação externos ao sistema. Ao recorrer a este género de serviços, o modelo passa a dispor de uma cobertura de nível internacional, beneficiando também da capacidade de evolução de alguns deles, que permitem a introdução de actualizações por parte de toda a comunidade.
As fontes consultadas não têm de ser formalmente Sistemas de Informação Geográfica. Qualquer fonte que permita a pesquisa sobre a localização de uma entidade pode contribuir para uma resposta. A Geo-Net-PT é uma fonte estruturada com dados geográficos administrativos de Portugal. É uma ontologia geográfica, desenvolvida na Faculdade Ciências da Universidade de Lisboa [Cha09] e disponível para a comunidade científica. Através deste recurso podemos descobrir, por exemplo, em que concelho ou distrito se situa uma localidade.

3.1.5 Tipos de pergunta

Como o sistema não está limitado a um domínio em particular, as perguntas formuladas podem incidir sobre qualquer assunto. A análise inicial à questão tem como objectivo apurar a natureza da mesma, para de seguida orientar a procura da resposta. Responder

a uma definição e encontrar a data de um evento passado são processos que num humano originam consultas à memória em moldes distintos. Se a pessoa não sabe as respostas e tenta encontrá-las numa enciclopédia, por exemplo, os critérios para identificar a resposta à definição não são idênticos aos que caracterizam a informação da data do evento. Deste modo, também o sistema recorre a técnicas distintas em função do tipo de questão que processa.

Categorias atribuídas

A estrutura morfo-sintáctica do texto da pergunta determina a categoria que lhe é atribuída. Os elementos chave na classificação são o tipo de interrogativo[2], se existir, e o verbo. O restante texto é também considerado, pois caracteriza o foco da questão, um eventual objecto de acção ou um contexto. As categorias de pergunta contempladas por este modelo são enumeradas em seguida. Para cada classe de questão apresentam-se exemplos representativos, bem como aspectos de processamento que são intrínsecos à categoria.

- **Definição**

  A categoria *Definição* é atribuída às perguntas em que se procura a descrição de uma entidade ou o significado de um conceito. Seguem-se alguns exemplos:

  O que era o A6M Zero?
  O que é a gripe suína?
  O que significa pandemia?

  Para responder a estas questões, o sistema efectua buscas em textos pertencentes às colecções locais, documentos na *Web*, dicionários locais e também dicionários disponíveis na *Web*. O significado de uma expressão composta por várias palavras raramente se encontra através de dicionários. Na maioria dos casos, as respostas resultam da análise de frases sobre o conceito a definir. Em particular, os textos da colecção Wikipédia são uma importante fonte de definições e descrições.

- **Quem**

  Esta categoria representa as questões que associam uma entidade a uma descrição ou acção. Nuns casos temos a descrição ou acção e pretende-se encontrar a entidade que lhe está associada. Noutras situações temos o nome da entidade, que é o foco da questão, esperando-se como resposta uma descrição para a mesma.
  A categoria *Quem* subdivide-se nas categorias:

  – QuemSerNome

[2] Advérbios e pronomes interrogativos: *que*, *quem*, *quanto(s)*, *quê*, *qual*, *quais*, *onde*, *quando*.

Este é o tipo atribuído às questões em que se pretende uma descrição para uma entidade concreta, identificada pelo nome ou uma possível designação. Uma forma simples deste género de pergunta é:

Quem foi Álvaro de Campos?

Como resposta, serão consideradas expressões que caracterizem a entidade em causa. Muitas vezes, a resposta descreve a vida, a profissão ou os feitos relevantes de uma pessoa. No caso apresentado, a resposta deve indicar que Álvaro de Campos é um heterónimo de Fernando Pessoa.
Esta categoria tem algumas semelhanças com o tipo *Definição*, no modo de procurar respostas. A diferença fundamental reside no facto desta categoria prever um foco da questão com identidade própria, enquanto que nas definições o foco pode ser qualquer conceito, concreto ou abstracto.

– QuemSerDescrição

Nas perguntas com a categoria *QuemSerDescrição* o procedimento é inverso ao que se enunciou para a categoria anterior. Nestes casos a questão contém uma descrição e o objectivo é encontrar a entidade a que corresponde. A pergunta seguinte é um exemplo destes casos.

Quem é o pato mais rico do mundo?

O tratamento das questões com este tipo envolve a procura da descrição ou equivalente em textos e na *Web*. Esses textos são examinados para extracção de entidades que possam estar vinculadas à descrição (directamente através do verbo ser, com recurso a anáforas ou mediante outra simbologia da linguagem escrita).

– QuemSujeitoDeVerbo

As duas categorias anteriores têm por base o verbo ser ou expressões que afectam uma descrição a uma entidade. Se a pergunta visa encontrar uma entidade que é sujeito de outro verbo, então pertence à classe *QuemSujeitoDeVerbo*. Este é o tipo usado quando se procura o agente praticante de determinada acção, que é descrita pelo verbo e eventuais complementos. Dois exemplos de perguntas com esta categoria são:

Quem fabrica os foguetões Ariane?
Quem ganhou a NBA em 1995?

As entidades que surgirem como agentes da acção indicada, directa ou indirectamente, serão coleccionadas como possíveis respostas. A acção é procurada em textos de documentos que incluam o verbo da questão, um verbo equivalente ou ainda uma especialização daquele. Para além do verbo, se existirem complementos eles são também considerados, nomeadamente quando identificam o objecto da acção ou um contexto geográfico ou temporal. Na pergunta sobre o vencedor da *NBA* é necessário seleccionar respostas válidas para o

contexto temporal pretendido: o ano de 1995.

– QuemComplementoDeVerbo
Como o nome sugere, esta categoria complementa a anterior, representando as perguntas em que se pretende a entidade sobre a qual é realizada alguma acção. A ilustrar esta classe de questões temos o exemplo:

A quem pertence o Cisne Branco?

A pergunta pode assumir várias formas. Neste exemplo o sujeito *Cisne Branco* surge após o verbo. Uma estrutura frequente nesta categoria é *sujeito-verbo-interrogativo*. Por vezes, a questão é apresentada na forma passiva e tanto esta categoria como *QuemSujeitoDeVerbo* podem ser atribuídas, com diferença na forma verbal capturada (verbo auxiliar e verbo principal ou apenas o verbo principal).

- **Quê**

Esta categoria representa as questões em que se usa o pronome interrogativo *quê* para representar uma coisa ou entidade, que é alvo de acção descrita por um verbo transitivo, ou que esteja associada a outro conceito através de uma preposição. As questões abaixo são representativas deste tipo.

Vasco da Gama descobriu o quê?
O lehendakari é presidente de quê?

Nesta classe de perguntas faz-se uma procura nas diversas fontes por frases ou outros elementos de informação, eventualmente estruturada, que sejam relativos ao foco da questão. Na primeira questão o foco é *Vasco da Gama* e na segunda o foco é *lehendakari*. Se o foco surgir associado a uma acção compatível (no primeiro caso) ou a um complemento directo em concordância com a pergunta (no segundo caso), então será extraída como resposta uma eventual entidade ou descrição.

- **QueQual**

A categoria *QueQual* atribui-se às questões em que se procura uma resposta de determinado tipo não numérico, explicitamente designado. O texto da pergunta inclui uma expressão junto ao interrogativo que descreve a natureza do conceito, entidade ou valor que se pretende. Vejamos dois exemplos concretos:

Que rio nasce na Serra da Estrela?
Qual é a montanha mais alta do México?

Para o primeiro caso o sistema tem de procurar uma resposta que seja um *rio*. Depois deve ainda satisfazer o critério descrito no restante texto. A segunda questão visa encontrar uma montanha, que terá de ser também a mais alta do México.
O processamento desta classe de questões prevê um teste de compatibilidade entre as possíveis respostas e o tipo indicado na pergunta, sempre que possível. O grau

de compatibilidade vai influenciar o peso a atribuir à resposta, sendo determinado pela base de conhecimento do sistema, complementada ainda com pesquisas em ontologias, nos textos locais e na *Web*, para além do contexto semântico, sempre que existir.

- **Quando**

  As questões cuja resposta é uma época, uma data ou outra indicação temporal são classificadas com o tipo *Quando*. É o caso das duas perguntas:

  Quando ocorreu a Batalha de Torres Vedras?
  Em que ano é que Halle Berry venceu o Óscar?

  O interrogativo *quando* associado ao verbo *ocorrer*, na primeira questão, indicam ao sistema que se procura o contexto temporal do evento *Batalha de Torres Vedras*. A resposta pode ser uma data completa, apenas o ano, a estação do ano ou o dia da semana. No segundo caso fica estabelecido à partida que a resposta deve ser do tipo *ano*. Se o sistema encontrar frases com expressões temporais relativas, estas serão interpretadas no contexto temporal do documento fonte, obtendo-se deste modo uma resposta concreta. Concretamente, se um texto publicado em 2003 contiver uma asserção com "*Halle Berry venceu um óscar no ano passado*", então o sistema pode responder com o ano de 2002.

- **Quantidade**

  Algumas perguntas requerem uma resposta quantitativa. Nestes casos o tipo atribuído é *Quantidade*, para o qual podemos considerar os exemplos:

  Quantas esposas tinha Ngungunhane?
  Qual é a temperatura do zero absoluto?

  A primeira questão tem como foco a entidade *Ngungunhane*. Um dos métodos para encontrar a resposta consiste em analisar frases sobre o foco da questão e procurar um valor numérico associado a *esposas*. Os valores numéricos podem ser expressos de diversas formas: por extenso, com algarismos (com ou sem separador decimal ou separador dos milhares), ou através de uma combinação de algarismos e texto. O segundo exemplo é um pouco diferente. O foco da questão é *zero absoluto* e o valor numérico pretendido diz respeito à *temperatura*, que é uma grandeza escalar com unidade de medida. Para a resposta ser correcta, a unidade de medida deve fazer parte da resposta, juntamente com a quantidade. Esta regra é válida para todas as grandezas escalares conhecidas pelo sistema. Por exigir uma unidade de medida, a detecção de possíveis respostas num texto é facilitada.

Algumas perguntas sobre quantidades são mais frequentes e têm especificidades na detecção e análise de respostas. Assim, esta categoria admite ainda os seguintes subtipos:

- QuantidadeIdade
  Esta categoria representa as perguntas sobre a idade de alguém ou alguma coisa. Em geral, o sistema procura a idade no tempo presente para a entidade foco da questão. Para determinadas questões é necessário encontrar a idade relativa a determinada época ou acontecimento. Consideremos os três exemplos que se seguem:

  Quantos anos têm os Rolling Stones?
  Quantos anos tinha Eusébio a 21 de Dezembro de 2000?
  Com que idade é que o Mequinho foi campeão brasileiro de xadrez?

  A pergunta inicial tem uma estrutura simples, denotando que o objectivo é a actual idade, expressa em anos, daquele famoso grupo de *rock*. A resposta pode ser encontrada num documento actual, calculada através de pontos de referência como o nascimento ou início[3], ou ainda pelos anos decorridos (X) desde o momento em que se conhecia ou detectou uma idade (Y, o que permite responder $X + Y$).
  O segundo caso condiciona a resposta a uma data. Para além das eventuais asserções sobre a idade que provenham já do contexto temporal previsto para a resposta, cada valor captado noutro contexto pode ser também projectado para o tempo da resposta. Se o sistema detectar uma idade válida para o ano 2005, tem de subtrair cinco anos para enquadrar a resposta na época pretendida. Quando as datas em análise incluem dia e mês, por exemplo relativamente ao nascimento e ao contexto temporal da resposta, o dia exacto permite apurar com mais rigor se passou ou não mais um aniversário.
  Na terceira questão a resposta deve inserir-se no mesmo contexto temporal do evento *ser campeão brasileiro de xadrez*, relativamente ao foco da questão, que é a entidade *Mequinho*. Isto obriga a um passo adicional para obter a data do evento e assim situar a resposta, como descrito para a segunda pergunta. Outra técnica usada é a extracção directa da resposta se o sistema encontrar uma frase com o evento sobre o foco da questão que contenham também um valor para a idade.

- QuantidadeDistância
  Esta categoria é atribuída às perguntas relativas a um comprimento, largura ou outra distância. Algumas perguntas do tipo *QuantidadeDistância* são:

  Qual o comprimento da Ponte do Øresund?
  Quantos metros saltou Nelson Évora em Pequim?

  A resposta é formada por um valor numérico e uma unidade de medida. Tal como para outras quantidades, a parte numérica pode ser expressa de várias

[3] Acções relevantes para cálculo da idade: *iniciar*, *ser inaugurado*, *ser fundado*, *nascer* (e equivalentes)

formas, pelo que o sistema normaliza o valor captado para uma representação com dígitos, sempre que possível. As unidades possíveis são as previstas no sistema métrico decimal para a grandeza comprimento, juntamente com unidades dos países anglo-saxónicos, dado que muitos os documentos na *Web* têm origem americana e inglesa.
O primeiro exemplo ilustra a versão mais simples deste género de perguntas, onde se pede a medida de determinada estrutura ou coisa. No segundo caso a distância pede-se explicitamente em metros. Para além disso, a resposta deve estar inserida num contexto geográfico relacionado com a cidade de *Pequim*.

– QuantidadePercentagem

Quando a quantidade pedida representa uma percentagem o sistema classifica a interrogação com o tipo *QuantidadePercentagem*. São perguntas em que se pretende determinar a fracção relativa de uma parte constituinte de alguma coisa maior. Os conceitos a que estas questões se podem aplicar estendem-se pelos mais variados domínios do saber. Como exemplo apresentamos as seguintes questões:

Qual é a percentagem de água no corpo humano?
Qual é a percentagem de finlandeses que fala sueco?

Tipicamente, a resposta encontra-se em frases sobre os conceitos global e constituinte. O valor numérico, textual ou não, surge muitas vezes associado ao símbolo %, à expressão *por cento* ou uma sua abreviatura. Noutros casos o documento fonte pode conter uma taxa relativa, de onde também se pode deduzir a percentagem.

– QuantidadePeso

Tal como as distâncias, o peso é um atributo solicitado de forma recorrente. Duas questões da classe *QuantidadePeso* são:

Quanto pesa o Airbus A380?
Quantos gramas tem um beija-flor?

As perguntas com este tipo nem sempre incluem a palavra *peso* ou uma conjugação do verbo *pesar*. O segundo exemplo mostra isso mesmo. A questão requer uma quantidade relativa à entidade *beija-flor*. É a unidade pedida para essa quantidade que permite classificar a questão com esta categoria.

– QuantidadePreço

As quantidades monetárias podem ser expressas de diversas formas. Não só a parte numérica pode apresentar várias representações, como também a moeda pode surgir pode extenso, de forma abreviada, através um símbolo como sufixo ou até como prefixo. Assim, quando a resposta pretendida é um valor monetário, o tipo atribuído é *QuantidadePreço*. As perguntas seguintes são exemplo desta categoria:

Qual foi o preço da Ponte Vasco da Gama?
Quantos dólares é que a Mars Observer custou?

Na forma mais comum desta categoria, a pergunta requer o *preço* ou *custo* de alguma coisa. Noutros casos é pedida uma quantidade com determinada unidade monetária, em geral associada a algum verbo conotado com valor pecuniário, como acontece no segundo exemplo.

– QuantidadeTempo

As questões sobre a amplitude de um intervalo de tempo, normalmente associado a um evento ou acção, são classificadas com o tipo *QuantidadeTempo*. Seguem-se dois exemplos desta categoria:

Quantos anos durou a Segunda Guerra Mundial?
Há quanto tempo faleceu João Paulo II?

Em geral, o que se pretende é determinar uma quantidade de tempo. A pergunta pode referir explicitamente a precisão para essa medida, ao especificar uma unidade de tempo para a resposta. Isto acontece no primeiro exemplo, onde se pede um valor em *anos*. A resposta a estes casos pode obter-se de dois modos. O modo mais simples é uma tentativa de encontrar o valor exacto directamente na análise de um texto. O modo indirecto envolve a detecção dos instantes inicial e final do acontecimento, apurando-se a resposta com o cálculo do tempo entre esses instantes.
Existem ainda situações em que o objectivo é encontrar o tempo decorrido desde determinada época ou acontecimento. É o caso do segundo exemplo, onde a resposta é o tempo desde determinada data até ao presente.
Alguns procedimentos como a detecção de datas em textos ou o cálculo de tempo decorrido entre dois instantes realizam-se de forma idêntica à praticada para as idades.

- **Localização**

Quando a resposta envolve uma análise geográfica ou posicional, a classe atribuída é *Localização*. É um tipo de perguntas muito usual, visando encontrar um lugar com determinado enquadramento ou apurar o local onde se situa determinada entidade ou em que decorre um evento. As interrogações seguintes são exemplo desta categoria:

Onde fica a Torre do Tombo?
Onde decorreu a Expo98?
Em que distrito fica Arraiolos?

A resposta à localização de um monumento, por exemplo, deve incluir uma cidade e um país, sempre que possível. Se aquilo que o sistema vai localizar é já o nome de uma cidade ou povoação, então a resposta deve conter uma eventual região e o país em se insere.

Neste género de perguntas o sistema recorre a bases de dados locais, juntamente com serviços de informação geográfica disponíveis via *Web*, para obter informações sobre cidades, países e locais relevantes. Estas são fontes de informação estruturada, de onde se extraem indicações precisas e que são complementadas com a análise de textos sobre o foco da questão.
Para além destas questões com estrutura relativamente simples, há outras em que o local a encontrar é descrito de modo mais complexo. As categorias seguintes são especializações de *Localização* vocacionadas para pesquisa com determinado tipo de critérios.

- LocalEmRegiãoComCritério

  Nestas questões pretende-se determinar que local corresponde a determinado critério genérico, obedecendo também a uma regra de localização. A questão determina um contexto espacial no qual se deve inserir a resposta. Consideremos dois exemplos concretos:

  Em que estado brasileiro fica Faro?
  Em que zona de Espanha nasce o rio Douro?

  A resposta à primeira pergunta deve ser um estado situado no Brasil. Ao pesquisar pela localização da povoação *Faro*, o sistema deve considerar apenas os resultados do Brasil, o que permite descobrir o estado do *Pará*.
  No segundo caso, o critério geral para o local a encontrar é ter a nascente do rio Douro. A restrição geográfica para o próprio local é pertencer ao país Espanha. Esta restrição funciona também como pista para o sistema.

- LocalizaçãoDeAcontecimentoRelativoAEntidade

  Esta categoria é atribuída às questões que requerem o local que é palco de algum acontecimento ou acção realizada sobre uma entidade. Nestas questões, ao contrário do previsto para a classe *Localização* geral, o verbo principal associado à entidade não pode ser *ficar*, *ser*, *decorrer*, *estar*, *localizar* ou equivalentes. Os exemplos seguintes ilustram a estrutura deste tipo de perguntas:

  Onde é que Joana de Arc foi queimada?
  Onde é que José Eduardo Pinto da Costa nasceu?

  Estes exemplos não incluem pistas para procurar directamente em sistemas de informação geográfica. É essencialmente o processamento do texto da pergunta e a análise de documentos, locais ou via *Web*, que conduzem a possíveis respostas.
  Para a primeira pergunta é necessário pesquisar documentos sobre a entidade *Joana de Arc*. Algum parágrafo sobre a sua morte na fogueira incluirá o nome da cidade francesa de *Rouen*, que será detectado na análise textual e validado como local pertencendo a França, constituído assim como resposta.
  No segundo exemplo, o sistema começa por pesquisar nos documentos relativos à entidade foco da questão, *José Eduardo Pinto da Costa*, recolhendo

como resposta locais associados ao seu nascimento.

As respostas a este género de perguntas nem sempre correspondem a países, cidades ou outras localizações geográficas. Existem situações cuja resposta pode ser uma posição relativa a outra entidade, um evento ou uma actividade. Concretamente, para o local onde *Joana de Arc* morreu poder-se-ia aceitar a resposta *"na fogueira"*.

- **Confirmação**

  Há questões cujo objectivo é comprovar uma afirmação ou verificar a veracidade de um facto. Normalmente, estas questões não incluem pronomes interrogativos. Um exemplo para esta categoria é:

  José Saramago venceu o Prémio Nobel da Literatura ?

  Nestes casos o sistema procura evidências sobre a afirmação. Se encontrar uma confirmação então a resposta é *sim*. Caso contrário, tenta ainda procurar evidências da negação da expressão. Para o exemplo apresentado, bastaria uma frase em que o mesmo verbo surgisse negado, relativamente às mesmas entidades (*José Saramago* e *Prémio Nobel da Literatura*). Havendo matéria para sustentar a negação da pergunta, a resposta é *não*. Quando não se encontram elementos para aceitar ou refutar a expressão, então a resposta do sistema é *não sei*.

- **TipoPorOmissão**

  Para outro género de perguntas ou para aquelas cuja estrutura não é entendida pelo classificador, a categoria atribuída é *TipoPorOmissão*. Esta é uma categoria vaga, usada nos casos em que menos se sabe sobre o que procurar como resposta. Ao processar uma questão deste tipo, o sistema procura frases com correspondência directa ou muito aproximada com o texto na questão. Daí, extrai-se como resposta qualquer expressão que nessa correspondência mapeie com um interrogativo.

O sistema procura respostas recorrendo a múltiplas estratégias, com procedimentos especializados para cada tipo de pergunta. Quanto mais concreta for a categoria atribuída à questão, maior será a certeza sobre o tipo de resposta a procurar.

Variantes na atribuição da categoria

A interpretação de uma pergunta nem sempre é óbvia. A expressão que descreve o foco da questão pode ser composta por várias palavras, o que dificulta a análise automática da pergunta. O mesmo se passa relativamente ao eventual objecto de uma acção ou complemento de um verbo no texto da questão. Consideremos a pergunta:

A Torre do Tombo foi construída há quanto tempo?

O foco da questão é determinado pela expressão *"Torre do Tombo"*. Esta expressão, por si só, pode conduzir o sistema a duas leituras possíveis para a pergunta. Tomando a expressão como um só nome, a pergunta é relativa ao arquivo central do Estado Português, situado em Lisboa. Mas existe outra interpretação possível. Se a leitura dos nomes que constituem aquela expressão se fizer de modo separado, o que é plausível neste caso, então a pergunta passa a incidir sobre uma torre situada em *Tombo*, uma localidade no estado de Pernambuco, no Brasil.
Para suportar esta análise, o sistema recorre a métodos de reconhecimento de entidades mencionadas, que são aplicados às diversas sequências de termos. As diferentes possibilidades para a interpretação de uma pergunta são consideradas pelo sistema. Sempre que a questão tem duas categorias possíveis ou quando a categoria atribuída admite vários parâmetros, nomeadamente para o foco da questão ou para um complemento, todas as variantes são exploradas e as suas respostas consideradas para a pergunta.

Extensibilidade do classificador

Ao disponibilizar o serviço a um conjunto de utilizadores, é comum identificarem-se padrões no género de perguntas. Por vezes, há perguntas que são apresentadas com uma estrutura frásica invulgar e que o sistema não classifica da melhor forma. Dada a complexidade da língua natural, isto pode acontecer com alguma frequência. Para evitar a realização de alterações sistemáticas ao código fonte do sistema, os critérios para a classificação de uma questão em determinada categoria podem ser externos ao software. O modelo contempla a extensibilidade do classificador mediante a leitura de um conjunto de regras que pode ser adaptado e estendido ao longo do tempo. Estas regras incluem expressões regulares, verificação de sinónimos e outros critérios, e são representadas com a norma XML.
Algumas estratégias de procura de respostas são baseadas na correspondência directa entre uma frase de um documento e uma expressão regular. Essa expressão, na sua versão mais simples, é composta pelo foco ou objecto da questão, um verbo e uma componente que vai mapear na possível resposta. Para os tipos de pergunta em que tal é possível, as regras com estas expressões regulares podem também ser externas à aplicação, facilitando o ajuste do sistema.

3.1.6 Espaço multidimensional de respostas

Para cada tipo de pergunta existem métodos específicos de procura de respostas candidatas. Destas, nem todas serão válidas. Mas para confirmar ou descartar uma possível resposta levando em consideração as demais, é necessária uma representação global que facilite a comparação de aspectos determinantes para essa decisão.
Cada resposta consiste numa expressão para o resultado, acompanhada de elementos sobre o contexto onde foi encontrada e ainda uma pontuação que reflecte o grau de confiança que o sistema lhe atribui. Depois de coleccionadas pelos diversos processos, as

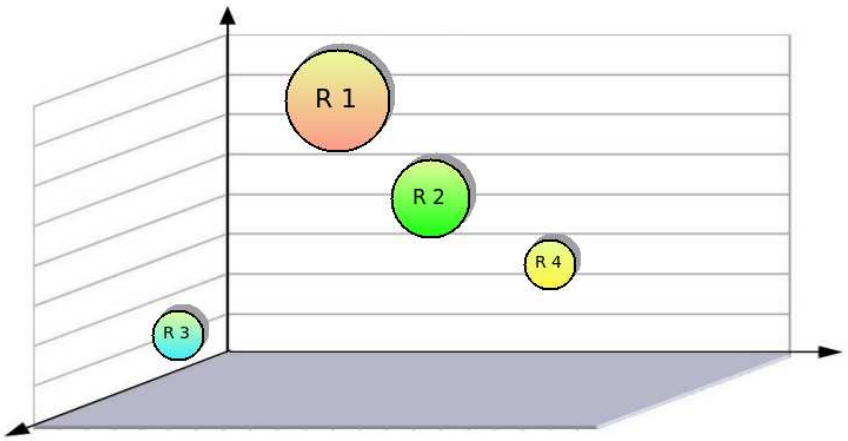

Figura 3.2: Disposição de respostas candidatas num espaço multidimensional

possíveis respostas são representadas num espaço multidimensional, como se ilustra na figura 3.2. O posicionamento de cada resposta candidata é determinado pelo seu valor e pelos atributos que lhe estão associados.

O conjunto de respostas forma assim uma nuvem de pontos. Em termos abstractos, se esses pontos estiverem aglomerados, isso significa que as respostas encontradas são próximas ou parecidas. Uma nuvem de pontos dispersos reflecte um conjunto com respostas díspares. Para cada uma das dimensões desde espaço de resultados, as coordenadas de uma resposta dependem do valor, do contexto e do tipo dessa resposta. Há três eixos fundamentais neste espaço de respostas, descritos em seguida, do mais simples para o mais complexo.

Eixo do Tempo

Sempre que se consegue uma referência temporal associada a uma resposta, essa informação vai assinalar a época em que se crê que a resposta é válida. Com um eixo que reflecte uma dimensão temporal, o espaço de respostas fica dotado de uma organização cronológica. Consideremos a seguinte questão:

O que está no centro do sistema solar?

A pergunta admite várias respostas, nomeadamente as decorrentes das teorias do geocentrismo e heliocentrismo. A figura 3.3 ilustra a disposição das respostas ao longo deste eixo. Esta organização do espaço de resultados facilita a escolha da resposta mais recente. Se uma pergunta incluir uma restrição temporal, então a intersecção entre o contexto temporal dessa questão e este eixo permite seleccionar as respostas válidas. Assim aconteceria se reformulássemos a pergunta para o ano 161 d.C., o ano em que

faleceu Ptolomeu[4], o que levaria à resposta *Terra* como centro do sistema solar.

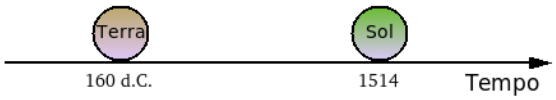

Figura 3.3: Respostas ao longo do eixo do tempo

Eixo do Espaço

Este eixo está associado à dimensão espacial, sendo um pouco mais complexo que o anterior por ser uma abstracção para uma localização, que pode ser representada de modo concreto, por coordenadas para latitude e longitude, ou de um modo simplificado com o nome da região geográfica, cidade ou país. Vejamos como seriam representadas as respostas da pergunta seguinte:

Quem é o presidente?

Considerando apenas algumas respostas actuais sobre a presidência de países, os resultados para Estados Unidos da América, Brasil, Portugal, França e Itália estão representados na figura 3.4. Todas as respostas são válidas para o respectivo país, devidamente assinalado neste eixo do espaço global de resultados. Se a pergunta não especifica um local em particular, a melhor resposta do sistema será aquela que, nesta dimensão, for mais compatível com o contexto espacial do próprio sistema (Évora, Portugal).

Eixo Semântico

Este é o eixo mais complexo do espaço de resultados. Na representação das respostas candidatas, esta dimensão constitui uma abstracção para a distância semântica entre os valores desses resultados. É estabelecida uma teia entre cada resposta e alguns conceitos relacionados, formando uma rede semântica. As ligações dessa rede vão permitir apurar o nível de semelhança entre os resultados.
A análise das respostas para este eixo é executada em função do tipo de pergunta e do tipo de valor encontrado para cada resposta.

- **datas**

  Importa esclarecer que uma resposta do tipo data pode ter associado um contexto temporal de valor diferente. Neste eixo, a análise incide sobre o valor da resposta

[4]Claudius Ptolomeu (83-161 d.C.) foi um matemático e astrónomo grego associado ao Geocentrismo. Antes de Ptolomeu, já outros defendiam que a Terra estava no centro do universo, incluindo Aristóteles.

Figura 3.4: Respostas distribuídas pelo espaço

e não sobre o contexto temporal em que foi encontrada. Tomemos como exemplo a questão:

Quando nasceu Thomas Mann?

A figura 3.5 inclui algumas das respostas candidatas que o sistema detecta em diversas fontes. Essas respostas são representadas de acordo com o seu significado. O valor *1702* representa um ano que vem antes do ano *1875*. A resposta *1875-06-06* representa um dia enquadrado no mês *Junho de 1875*, que por sua vez se insere no ano *1875*. A representação usada reflecte relação entre as respostas pelo significado que têm.

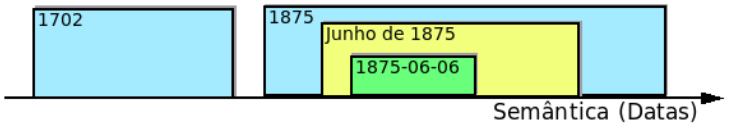

Figura 3.5: Dimensão semântica: representação de datas

- **quantidades**

  As respostas candidatas com quantidades podem surgir com formatos distintos, no que respeita a separador decimal ou separador dos milhares. Um valor numérico inteiro pode ser escrito por extenso, com algarismos ou de modo misto. Ao colocar cada resposta neste eixo, o sistema interpreta o significado do valor, no formato encontrado, e representa-o de um modo normalizado. Se existirem unidades associadas ao valor numérico, esta interpretação faz-se para cada unidade em separado,

a não ser que o sistema consiga efectuar uma conversão entre as unidades. Para a pergunta:

Quanto pesa um beija-flor?

o sistema sabe que os símbolos *g* e *gr* são usados para a unidade *gramas*. Assim, os valores numéricos com estas unidades ficam representados num mesmo eixo, como mostra a figura 3.6.

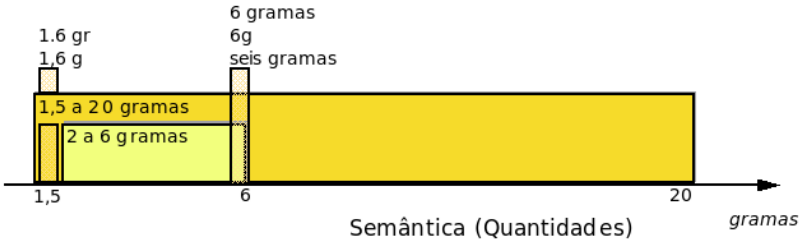

Figura 3.6: Dimensão semântica: representação de quantidades

Os resultados *seis gramas* e *6g* são equivalentes, ficando portanto no mesmo ponto deste eixo. O mesmo acontece com as respostas *1.6 gr* e *1,6 g*. As outras respostas assinaladas nesta dimensão do espaço de resultados são intervalos. O intervalo mais amplo é entre *1,5 a 20 gramas* e outro está contido no primeiro, entre *2 a 6 gramas*. A variedade de respostas resulta das diferentes proveniências destes valores.

- **factóides**

  As respostas deste género são usualmente formadas por uma única palavra ou uma expressão curta correspondente ao nome de uma entidade. A disposição destes factóides na dimensão semântica do espaço de resultados faz-se através de uma malha de conceitos interligados. O objectivo é representar o significado das palavras por forma a evidenciar semelhanças entre as respostas, no campo semântico. Estabelecendo um paralelo com o processo neurocognitivo de activação de associações entre conceitos, sobre as memórias semântica e episódica [DMP07], cada factóide é associado a um conjunto de termos relacionados, que não têm de fazer parte da pergunta ou do documento onde é encontrada a resposta. Este processo é genérico, independente de qualquer domínio em particular e teoricamente funcional para todo o léxico de qualquer idioma. O procedimento adoptado é inspirado em trabalhos anteriores sobre representação semântica e detecção de similaridade, de Kozima [KF93], e desambiguação de Veronis[VI90] e Tsatsaronis [TVA07], com base em redes associativas e dicionários.
  A fase inicial consiste na representação das respostas candidatas no espaço, ainda sem diferenciação de posicionamento entre elas, mas com indicação da pontuação

de cada uma, atribuída durante o processo de extracção de respostas. Consideremos a seguinte questão:

Que animal é o Cocas?

A resposta esperada é um factóide. A figura 3.7 mostra algumas das respostas candidatas que o sistema detecta para esta questão. Os factóides mostrados são respostas plausíveis encontradas em textos. Apesar da interpretação mais comum da pergunta nos remeter para a famosa personagem da série *Os Marretas*, é possível que algum outro animal com o mesmo nome seja referido nos recursos consultados, o que conduz a outras respostas válidas, apesar de menos populares.

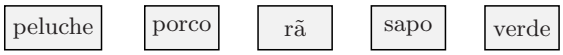

Figura 3.7: Dimensão semântica: factóides (*início*)

O passo seguinte é a obtenção do primeiro nível de informação semântica para as possíveis respostas. A cada nó resposta é adicionado um elo de ligação para novos nós que contém possíveis significados ou conceitos relacionados. Os elos de ligação têm um tipo e um peso associados. O tipo caracteriza a relação semântica expressa pelo arco entre os conceitos dos nós de origem e destino. O peso reflecte o grau de certeza do sistema sobre essa relação. O peso de uma relação é um valor real p_{rel}, tal que $0 < p_{rel} <= 1$, que irá influenciar o processo de propagação do nível de activação, conforme será explicado na secção seguinte.
Para cada resposta desta categoria, tipicamente formada por uma palavra, importa coleccionar alguns conceitos que com ela estabelecem uma relação semântica. A procura dessas relações semânticas faz-se pela definição da palavra (ou da respectiva forma base) num dicionário acessível via *Web*. Com base na metodologia seguida pelo sistema SESEMI[5] [Van95], mas de modo simplificado, o sistema processa a análise sintáctica de uma definição com um conjunto de regras que validam a existência ou não de determinado padrão. Essas regras são focadas em nomes, verbos e adjectivos presentes na estrutura sintáctica que evidenciem relações como hiperónimo, merónimo, sinónimo, antónimo e agente de determinado verbo, quando possível.
O quadro seguinte tem uma lista com as definições dos termos da figura 3.7.

[5]SESEMI: *System for Extracting SEMantic Information*. Identifica relações semânticas através de consultas a dicionários *online*. As definições nas entradas dos dicionários têm determinada estrutura que facilita a detecção de padrões que podem evidenciar relações semânticas [Van95]. Em trabalho mais recente e para o Português, foi usada uma abordagem análoga para a construção de uma ontologia com base na extracção de informação semântica de dicionários [OSGS08].

peluche *s.m.* - Boneco revestido com tecido felpudo.
porco *s.m.* - Quadrúpede da classe dos mamíferos e da família dos suídeos.
adj. - Imundo
rã *s.f.* - Batráquio saltador e nadador, de pele lisa, verde-parda.
sapo *s.m.* - Espécie de batráquio anuro.
verde *s.m.* - A cor verde.

As definições com adjectivos indicam um significado possível para o termo, como acontece com *porco*, e são captadas como sinónimos. As entradas do dicionário que descrevem substantivos são mais extensas, pelo que se usa um analisador morfo-sintáctico. A figura 3.8 mostra a estrutura sintáctica obtida para a definição de *peluche*. O padrão existente leva à associação entre o conceito *peluche* e o elemento nuclear da estrutura, o substantivo *boneco*.

```
NPHR:np
=H:n('boneco' <H> M S)                          boneco
=N<:v-pcp('revestir' M S)                       revestido
=N<:pp
==H:prp('com')                                  com
==P<:np
===H:n('tecido' <cc-rag> <mat-cloth> M S)       tecido
===N<:adj('felpudo' M S)                        felpudo
```

Figura 3.8: Dimensão semântica: estrutura sintáctica da definição

O uso do dicionário para captar significados para as respostas candidatas apresentadas permite aumentar o espaço de resultados com novos nós e relações, designadamente hiperónimo e sinónimo, como se ilustra na figura 3.9.

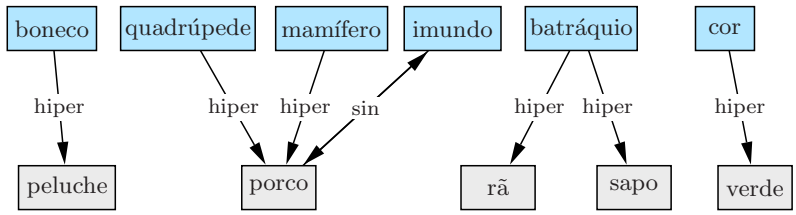

Figura 3.9: Dimensão semântica: factóides (*1º nível de relações*)

Os factóides com nomes de entidades são mais difíceis de caracterizar neste primeiro nível de expansão, uma vez que não possuem uma definição de dicionário. Nesses casos determina-se simplesmente a categoria da entidade referida, que pode ser uma pessoa, um monumento, uma empresa ou outra classe. Esta informação está contida nos catálogos de entidades. A relação semântica estabelecida entre a

entidade e a sua categoria é '*instância de*'.
Após a consulta de dicionários e catálogos de entidades, a malha de conceitos é agora constituída pelos nós com as respostas e mais um nível de nós com os conceitos semanticamente relacionados. O terceiro passo é a expansão da malha por importação de termos relacionados com cada nó em estruturas semânticas associativas acessíveis pelo sistema. A expressão em cada nó é procurada em ontologias e em redes semânticas e os respectivos conceitos relacionados são importados para o espaço de resultados. Os novos nós terão um elo de ligação com peso 1, estabelecido de acordo com a estrutura de origem.
A rede semântica resultante deste processo está indicada na figura 3.10.

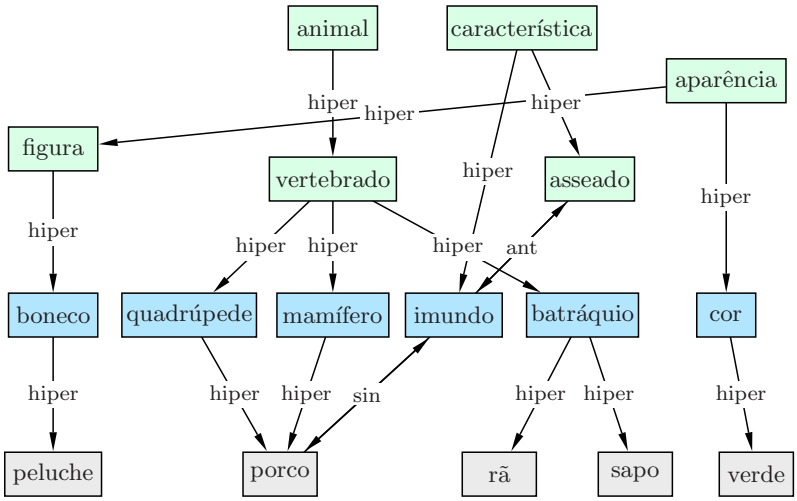

Figura 3.10: Dimensão semântica: factóides (*rede de conceitos*)

- **frases**

  As respostas mais compridas estão associadas a perguntas do tipo *Definição*. É o caso da pergunta:

  O que é um berimbau?

  As respostas candidatas com maior pontuação que o sistema encontrada nos documentos são as seguintes:

uma harpa de corda única
um instrumento de percussão usado tradicionalmente na capoeira
um instrumento musical afro-brasileiro
um elemento fundamental na capoeira
uma arma maligna e mortal
feito de um bastão de madeira

  A transposição destas expressões para a dimensão semântica do espaço de resultados consiste na identificação dos termos relevantes da frase e na percepção dos seus possíveis significados. A análise morfo-sintáctica do texto de cada resposta

gera uma estrutura através da qual se detectam os termos de núcleo. A figura 3.11 mostra a estrutura sintáctica para a primeira resposta do exemplo. As etiquetas do analisador com a cor azul denotam núcleo (H), substantivo (n) e adjectivo (adj). Os conceitos núcleo da expressão surgem destacados a vermelho. O adjectivo *único* é também captado como característica do nome *corda*, que por sua vez é um núcleo secundário, parte constituinte do conceito definido. Os termos

```
NPHR:np
=>N:art('um' <arti> F S)        uma
=H:n('harpa' <tool-mus> F S)    harpa
=N<:pp
==H:prp('de')                   de
==P<:np
===H:n('corda' F S)             corda
===N<:adj('único' F S)          única
```

Figura 3.11: Dimensão semântica: estrutura sintáctica e elementos de núcleo

núcleo são representados em nós, numa rede semântica, sendo considerados nós de referência da resposta. Cada nó de referência é relacionado com outros nós do espaço através de relações semânticas, aplicando-se-lhes o procedimento já descrito para a representação dos factóides. Uma resposta é simbolizada por um nó associado aos seus nós de referência através de relações que dependem da estrutura sintáctica da expressão, como ilustrado na figura 3.12. O nó *'corda única'* é um nó derivado do nó de referência *corda*, simbolizando um conceito daquele tipo que é caracterizado ainda por outros elementos semânticos (neste caso apenas uma relação com o adjectivo *único*). Havendo um nó derivado, é com ele que o nó da resposta se associa, em vez de *corda*, estabelecendo com ele uma relação que depende da estrutura sintáctica. No caso, a etiqueta *mer* simboliza a relação merónimo, significando que *'corda única'* é parte do conceito naquela resposta.

Com todos os resultados neste espaço multidimensional, temos os respectivos significados no eixo semântico, e o espaço e tempo a que se referem nos restantes eixos. Podemos agora reavaliar o conjunto e eventualmente tender para uma das respostas. É disso que trata próxima a secção.

3.2 Promover, preterir, escolher

No momento em que uma resposta é extraída de um texto, a pontuação que lhe é associada depende em grande medida da distância ao foco da questão e da estrutura da frase de suporte. O modelo apresentado prevê uma segunda fase de avaliação da resposta, posterior à fase de extracção, onde o valor de cada resposta é analisado no âmbito

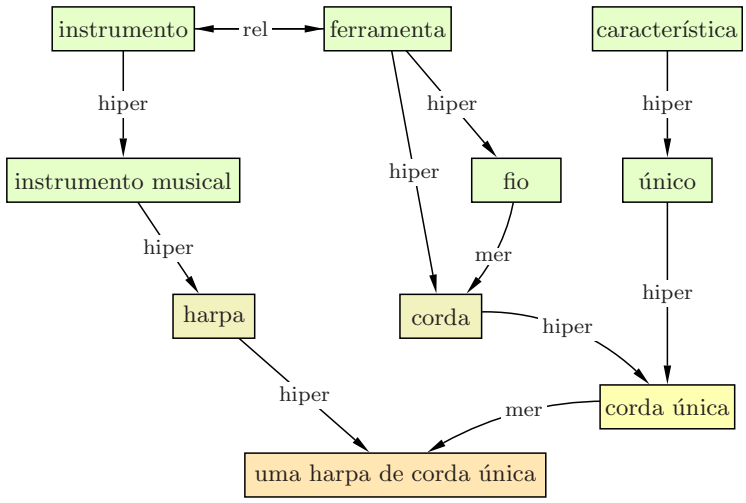

Figura 3.12: Dimensão semântica: representação de expressões

da pergunta, considerando também todas as respostas alternativas e demais informação disponível. Deste processo resulta uma nova ordenação das resposta candidatas, onde se espera que as respostas mais relevantes sejam as mais cotadas.

Recorrendo ao espaço multidimensional, o sistema deve valorizar respostas cujos contextos espacial, temporal e semântico sejam compatíveis com previsto para a pergunta. Deve dar-se preferência a resultados cujo contexto espacial, ou local a que estão vinculados, esteja de acordo com uma eventual restrição no texto da questão ou, na ausência de tal restrição, que seja admissível na óptica de quem responde, pela proximidade ao local de referência do sistema. O eixo do espaço torna mais fácil a automatização deste processo de escolha. Voltemos ao exemplo da figura 3.4 que, de modo simplificado, regista o país para cada resposta sobre os presidentes. Se a pergunta não especifica uma região, então o sistema escolhe o resultado da região em que se enquadra, Portugal, cujo presidente é *Cavaco Silva*.
No que respeita ao contexto temporal, o procedimento geral é dar prioridade às respostas cujo tempo de validade é mais próximo do momento presente. No exemplo anterior sobre o centro do sistema solar, o eixo do tempo na figura 3.3 facilita a identificação do resultado mais actual: o *Sol*. Mas se a pergunta nos remete para um ano ou época em particular, então os resultados mais próximos dessa referência é que devem ter primazia.

Naturalmente, o valor de cada resposta também é objecto de análise nesta fase de escolha. E há três tipos de abordagem sobre o valor das respostas:

- confirmação em diversas fontes - as respostas que são suportadas por várias fontes poderão ser encaradas como mais fiáveis.
- forma - duas respostas podem ter o mesmo significado mas apresentarem formatos sintácticos distintos, por exemplo nas definições. Se um desses formatos for mais

sucinto ou típico para o género de pergunta, então a respectiva resposta deve ser valorizada.

- significado - deve dar-se mais relevo às respostas cujo significado esteja relacionado com a pergunta ou com os interesses do utilizador. Pode procurar-se o significado comum à maioria das respostas e preferi-las em detrimento das restantes. A proximidade semântica entre um grupo de respostas pode servir para reforço de confiança naquela área do espaço de resultados. O eixo semântico em que as respostas foram representadas permite a sistematização destes procedimentos.

Em qualquer dos casos, as técnicas usadas são dependentes da categoria de pergunta, pois esta determina um conjunto de características esperadas para a resposta.

3.2.1 Validação e confirmação em diversas fontes

Algumas expressões recolhidas como possíveis resultados não constituem respostas válidas para a pergunta. A automatização de um processo que filtre essas falsas respostas contribuirá para a eficácia do sistema. Esta área tem sido explorada nos últimos anos, nomeadamente através de actividades como o AVE[6]. Neste desafio, os sistemas participantes fazem uma apreciação de várias respostas para cada pergunta e, tendo em conta também uma frase ou parágrafo de origem, deliberam sobre a validade ou não de cada possibilidade. O princípio subjacente é a implicação textual[7] entre o texto de suporte e a resposta candidata. Esta é combinada com a pergunta na formulação de uma hipótese que poderá ou não ser dedutível a partir dos elementos disponíveis [PRSV08]. Tomando como exemplo a questão

Quem é o Presidente da República?

e a resposta candidata *Cavaco Silva*, podemos construir uma hipótese com a forma afirmativa da pergunta e a resposta a avaliar:

Cavaco Silva é o Presidente da República.

Se a origem do resultado *Cavaco Silva* é uma expressão que implica a hipótese, isto é, uma expressão a partir da qual se conclui que a hipótese é verdadeira, então a resposta candidata é encarada como válida. Deste modo, a validação de respostas pode ser encarada como um problema de reconhecimento de implicação textual entre o texto de origem dos resultados e hipóteses que resultam da instanciação de uma forma padrão da pergunta [Yus08]. Estas formas padrão variam em função da categoria de pergunta.
Por outro lado, tendo encontrado uma resposta candidata em determinada fonte, o sistema pode verificar a plausibilidade desse resultado noutra fonte. Mais uma vez, procura-se tirar partido da redundância de informação na *Web*. Agora o objectivo não é detectar um valor para a resposta mas sim corroborar ou medir a popularidade de

[6] AVE: *Answer Validation Exercise*. É uma actividade para sistemas de validação de respostas [PVSV06] [PVV07] [VPV08], associada à tarefa de Pergunta-Resposta do *Cross Language Evaluation Forum*.

[7] *textual entailment*

uma resposta. Para isso, realiza-se uma pesquisa pelas hipóteses geradas para a resposta e de acordo com o tipo de questão. O excerto de código da figura 3.13 mostra as consultas para confirmação de cada resposta `tmp`, definidas pela categoria de pergunta, pelo método `getExpressoesPesquisaWeb()`. Se a pesquisa obtiver resultados, então existe texto de suporte para a hipótese, e consequentemente para a resposta candidata.
Alguns processos de validação obrigam a um reconhecimento de implicação textual indirecto e complexo. Em vez da verificação estatística baseada no grande volume de dados, recorre-se à inferência sobre uma base de conhecimento. Esta abordagem é mais usual em sistemas de domínio fechado [MNPT02b], com informação semântica suficientemente abrangente para a área temática respectiva. Um sistema de âmbito geral também pode beneficiar da existência de informação semântica, designadamente para a análise linguística da pergunta, para apreciar a compatibilidade entre cada possível resultado e o tipo de resposta esperado para a categoria da pergunta, ou ainda na verificação de implicação textual envolvendo termos semanticamente relacionados.
Após a fase de validação, o grau de confiança das respostas é ajustado através do aumento ou diminuição da respectiva a pontuação.

```
Vector<Answer> vresps= respostas0.sortResp();
int count= vresps.size();
int[] ocorrencias= new int[count];
for (int i=0; i<count; i++) {
    ocorrencias[i]= 0;  // reset
    Answer answer= vresps.get(i);
    String tmp= answer.getAnswer();
    Vector<String> vexps= categoriaQuestao.getExpressoesPesquisaWeb(tmp);
    // 1. conta as ocorrências para cada expressão
    for (int k=0; k < vexps.size(); k++) {
        String exprPesq= vexps.get(k);
        debug("\t PESQ i: "+i+"  -->"+exprPesq+"<--");

        SensoGoogleSearch g= new SensoGoogleSearch();
        GoogleSearchResult gsr= null;
        int oc= 0;
        try {
            gsr= g.pesquisaApenasContagem(exprPesq);
            oc= gsr.getTotal(); // resultados da pesquisa
        }
        catch (Exception e01) {
            e01.printStackTrace();
        }
        // --
        //debug("\t  web search (i="+i+",j="+k+") resultados: "+oc+"  == "+tmp);
        ocorrencias[i]+= oc;
    }
}
```

Figura 3.13: Código fonte: múltiplas consultas para comprovar respostas via Web

3.2.2 Análise por propagação de activação semântica

Comparar automaticamente os resultados ao longo do eixo semântico é simples para quantidades ou datas. Para as respostas com texto (factóides ou frases) a comparação não é tão imediata. Ao decidir entre duas possíveis respostas, um ser humano não observa apenas a morfologia das palavras em cada resposta, nem considera apenas as definições estritas dos respectivos termos. Uma pessoa analisa essas respostas tendo em conta também as recordações que cada termo suscita na sua memória e os conceitos que explícita ou implicitamente lhes estão associados.
Neste modelo, os significados possíveis para um termo estão expressos na rede de conceitos onde as respostas foram representadas. A detecção de semelhança semântica entre respostas textuais faz-se pela análise do espectro semântico dos respectivos nós na malha de resultados. A técnica usada tem inspiração em teorias sobre o funcionamento da memória humana, baseando-se no modelo de propagação de activação semântica sobre estruturas associativas [Cre97].
O espectro de activação semântica para uma resposta R obtém-se através da propagação do nível de activação ao longo dos nós da rede, de acordo com o seguinte:

- Após a construção do espaço de resultados e a fase de validação, cada nó de resposta tem nível de activação inicial igual à pontuação da resposta que representa. Os restantes nós da rede começam com nível de activação zero.

- Existe a noção de estado de um nó, que pode ser activo ou inactivo. O estado de um nó é activo se o seu nível de activação é maior ou igual ao limite de activação para o conjunto de resultados. Este limite é calculado em função da pontuação das melhores respostas candidatas. Corresponde a 80% da pontuação da vigésima resposta mais cotada até então, ou da última, caso haja menos de vinte resultados possíveis.

- No início do algoritmo, toma-se como activo apenas o nó relativo a R. Para os restantes nós resposta o nível de activação é ignorado. O processo começa com o nó da resposta R, cujo nível de activação é propagado pelo espaço semântico através dos nós adjacentes, como se descreve adiante.

- Após esta fase, o nível de activação de diversos nós terá sofrido uma eventual alteração. É formada uma nova lista com nós que estejam agora em estado activo e que não tenham sido processados antes. Realiza-se nova iteração sobre os nós desta lista. Se daí resultar a passagem de mais nós ao estado activo, então o processo continua. A figura 3.14 apresenta o código fonte do método `propagacaoDeActivacaoDeNo()`, que executa o procedimento descrito.

- A propagação da activação de um nó para outro adjacente depende da relação semântica que os une. Em geral, uma relação admite a propagação só num dos sentidos (para o mais geral), excepto para as relações *sinónimo*, *antónimo*, *relacionado* e *oMesmoQue*, que permitem propagação em qualquer dos sentidos. O nó de destino da propagação altera o seu nível de activação com base em $VProp$,

que depende de dois parâmetros da relação e da activação de referência na origem, como indica a equação 3.1. Este último valor corresponde ao próprio nível de activação, se a propagação tem origem no nó resposta base para o processo. Para os restantes casos, a activação de referência $ActRef_{origem}$ é o valor da última propagação recebida no nó de origem. O número total de propagações recebidas e os respectivos valores também são considerados, como mostra a equação 3.2.

$$VProp = ActRef_{origem} \times PR \times CP \tag{3.1}$$

$$NA = NA_{inicial} + \sum_{j=1}^{\#props} (K \times \frac{CP_j}{|CP_j|}) + \frac{\sum_{j=1}^{\#props} VProp_j}{\#props} \tag{3.2}$$

Na primeira equação, PR representa o peso da relação. O parâmetro CP é o coeficiente de propagação da relação. Trata-se de um valor entre -1 e 1. Um valor negativo significa que inibe a activação no nó de destino, por este ter um significado incompatível com o significado do nó de origem, como sucede com a relação *antónimo*. A tabela 3.1 mostra valores para o coeficiente de propagação do nível de activação para várias relações semânticas, numa configuração possível do modelo. Estes valores podem ajustar-se em função das características da rede semântica do espaço de resultados. A passagem de $VProp$ para os nós vizinhos é executada pelo método `propagacaoDeActivacaoParaVizinhos()`, apresentado na figura 3.15.
A equação 3.2 mostra como é determinado o nível de activação para um nó após ter sido alvo de *#props* propagações. Ao nível de activação inicial do nó é somado um ajuste formado por uma parcela proporcional ao número de propagações recebidas e outra parcela com a média dos valores recebidos nas mesmas propagações. A constante K corresponde a um valor inteiro que vai ser multiplicado na parcela proporcional à quantidade de propagações recebidas. A fracção de CP_j serve para multiplicar K por 1 ou por -1, em função do sentido do coeficiente de propagação usado na propagação índice *j*. Após alguns testes com o algoritmo, determinou-se o uso de $K = 5$ para o incremento gradual do nível de activação, sem aumento demasiado rápido ao longo da rede de nós. O método `commitValorDoNivelDeActivacaoAposPropagacao()` na figura 3.15 faz a actualização do nível de activação do nó, conforme foi descrito.

RELAÇÃO	CP
hiperónimo	0,8
merónimo	0,4
sinónimo	0,8
antónimo	-0,8
relacionado	0,4
instânciaDe	0,8
localizadoEm	0,8
oMesmoQue	0,8

Tabela 3.1: Coeficiente de Propagação do Nível de Activação

```
private void propagacaoDeActivacaoDeNo(NoDeEspaço no1, int marcaDeExecProp) {
    // 1: processa este no:
    no1.setNoBaseDaPropagacao(true);
    no1.propagacaoDeActivacaoParaVizinhos(marcaDeExecProp,this);
    commitValorDoNivelDeActivacao();
    // ----
    // 2: ver se ha alteracao: mais nos activos
    Vector<NoDeEspaço> novosActivos= null;
    do {
        novosActivos= new Vector<NoDeEspaço>();
        Vector<NoDeEspaço> vactivos= getNosActivosQueReceberamPropagacao();
        for (int i=0; i<vactivos.size(); i++) {  // todos os activos agora...
            if ( vactivos.get(i).aindaNaoExecPropagacaoAposMarca(marcaDeExecProp) )
                novosActivos.add(vactivos.get(i));
        }
        // ----
        // 3: com nós que AGORA estão ACTIVOS e que ainda não executaram propagação
        for (int i=0; i<novosActivos.size(); i++) {
            NoDeEspaço esteNoActivo= novosActivos.get(i);
            esteNoActivo.propagacaoDeActivacaoParaVizinhos(marcaDeExecProp,this);
            commitValorDoNivelDeActivacao();
        }
    } while (novosActivos.size()>0);
    // ---
    no1.setNoBaseDaPropagacao(false);
}
```

Figura 3.14: Código fonte: método para propagar o nível de activação de um nó (Classe EspaçoDeResultados)

```
public void propagacaoDeActivacaoParaVizinhos( int marcaDeExecPropagacao,
                                               EspaçoDeResultados espaco) {
    this.marcaCodigoDeExecPropagacao( marcaDeExecPropagacao );

    Iterator<RelacaoEspaço> iter = setRelacoesOrigem.iterator();
    while (iter.hasNext()) {
        RelacaoEspaço rel= iter.next();
        if (espaco.relacaoPropagaActivacaoOrigemDestino(rel)) {
            if (rel.getNoDestino().isNoBaseDaPropagacao())
                continue;
            // VProp = ActRef * PesoRelacao * CoeficientePropagRel
            int vprop= Math.round(rel.getNoOrigem().getNivelDeActivacaoLidoParaPropagar()
                                * rel.pesoRelacao() * rel.coeficientePropagacaoRel() );
            rel.getNoDestino().registaContribuicaoActivacao(vprop);
        }
    }

    iter = setRelacoesDestino.iterator();
    while (iter.hasNext()) {
        RelacaoEspaço rel= iter.next();
        if (espaco.relacaoPropagaActivacaoDestinoOrigem(rel)) {
            if (rel.getNoOrigem().isNoBaseDaPropagacao())
                continue;
            // VProp = ActRef * PesoRelacao * CoeficientePropagRel
            int vprop= Math.round(rel.getNoDestino().getNivelDeActivacaoLidoParaPropagar()
                                  * rel.pesoRelacao() * rel.coeficientePropagacaoRel() );
            rel.getNoOrigem().registaContribuicaoActivacao(vprop);
        }
    }
}


public void commitValorDoNivelDeActivacaoAposPropagacao() {
    int propagRecebidas= getNumeroDePropagacoesRecebidas();
    if (propagRecebidas>0) {  // se ha alteracoes, caso contrario nao faz nada
        int total= 0;
        for (int i=0; i<listaContribuicoesPropagActivacao.size(); i++)
            total= total + listaContribuicoesPropagActivacao.get(i).intValue();
        // --
        int media= total / propagRecebidas;
        int ajuste= somatorioKCPj() + media;
        nivelDeActivacao= _nivelDeActivacaoINICIAL + ajuste;
        _jaRecebeuPropagacao= true;
    }
}
```

Figura 3.15: Código fonte: propagação de um nó para os nós vizinhos (Classe NoDeEspaço)

O espectro de activação semântica de uma resposta R é constituído pelos segmentos da rede que relacionam R com os nós que através da propagação passaram ao estado activo.

O exemplo com as respostas candidatas listadas na figura 3.7 leva a um espaço de resultados onde o limite de activação é 65. O espectro obtido para cada uma dessas respostas está representado nas figuras 3.16 a 3.20. Os nós activos aparecem destacados com um contorno vermelho. Junto ao texto de cada nó é indicado o valor numérico com o nível de activação após a conclusão do algoritmo de propagação.
A primeira resposta resultou na activação de um só nó vizinho, uma vez que o nó *figura* ficou com nível 59, abaixo do limite de activação. Já na figura 3.17, o processo terminou com mais três nós em estado activo, devido ao facto do nó base da propagação ter mais vizinhos (dado que o seu nível de activação inicial era próximo ao do primeiro caso).

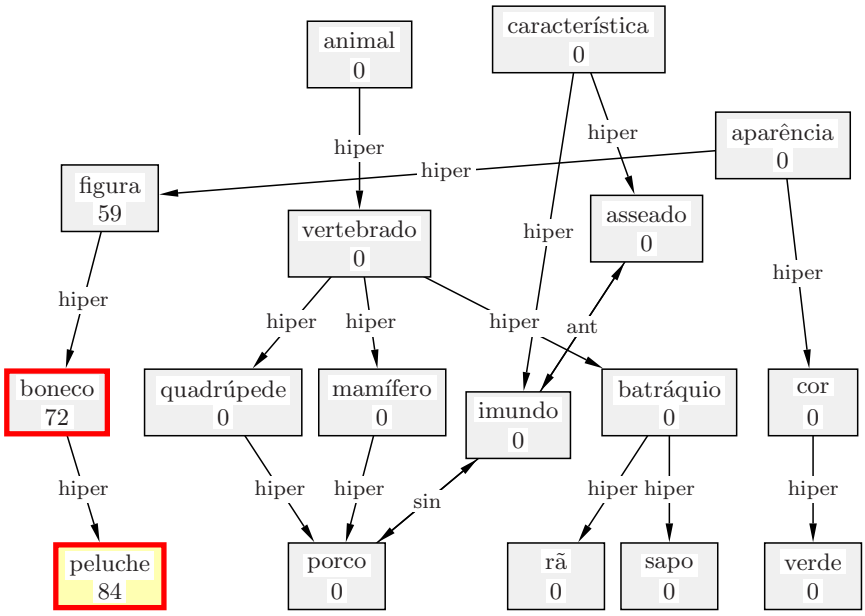

Figura 3.16: Espectro de activação semântica (resposta *peluche*)

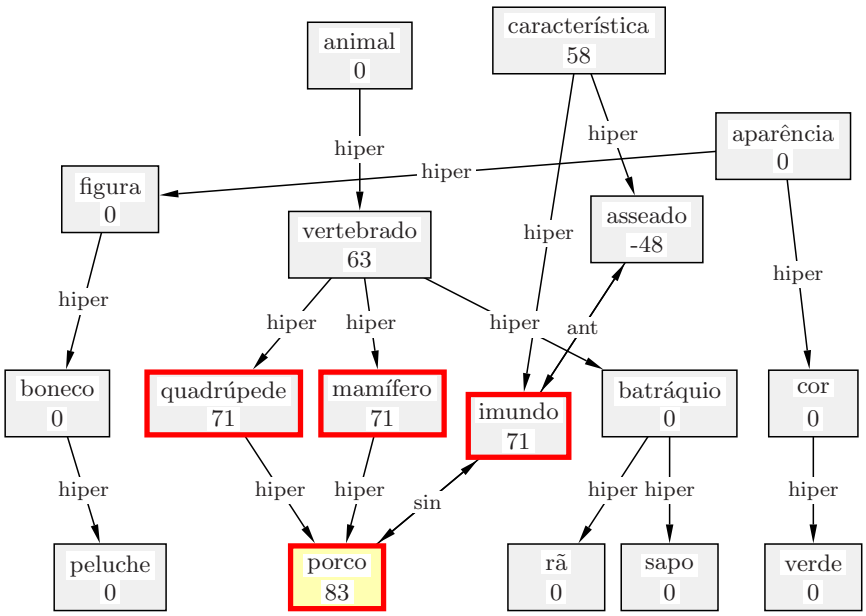

Figura 3.17: Espectro de activação semântica (resposta *porco*)

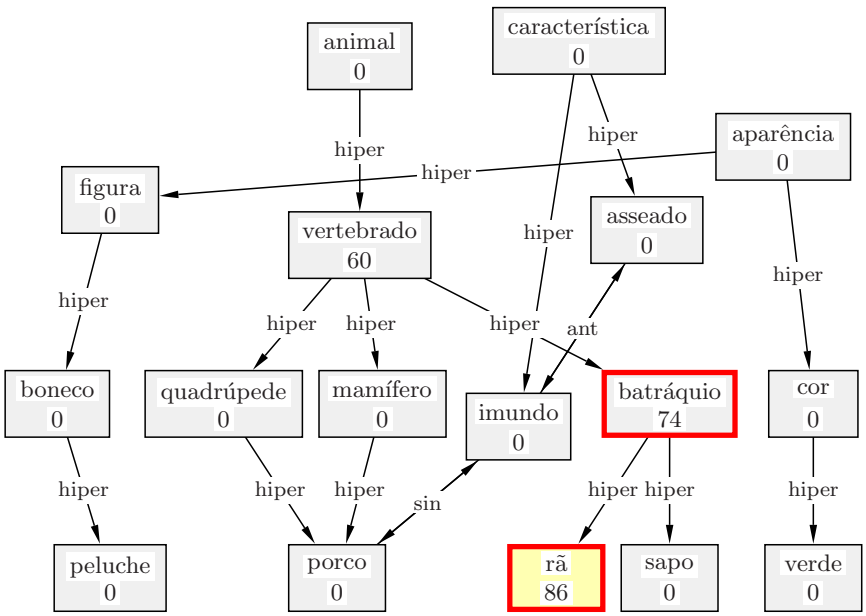

Figura 3.18: Espectro de activação semântica (resposta *rã*)

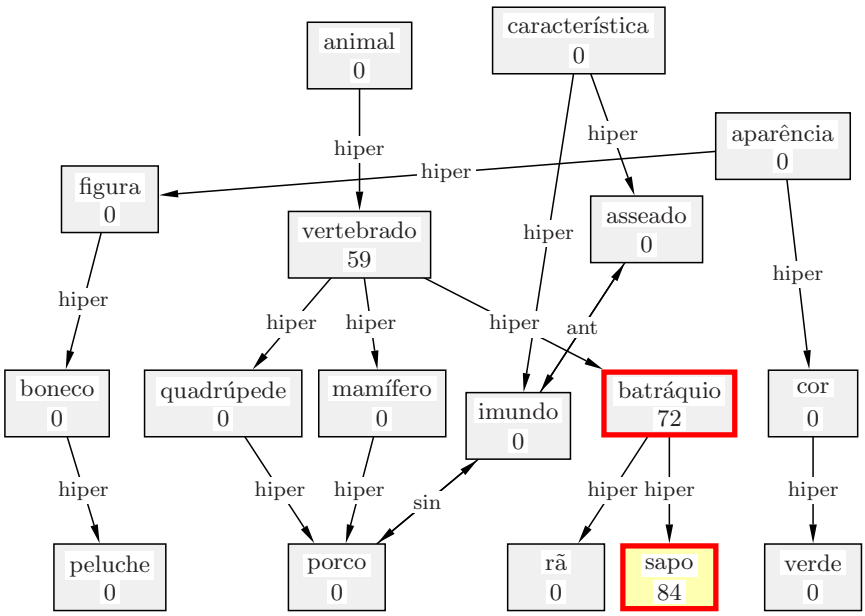

Figura 3.19: Espectro de activação semântica (resposta *sapo*)

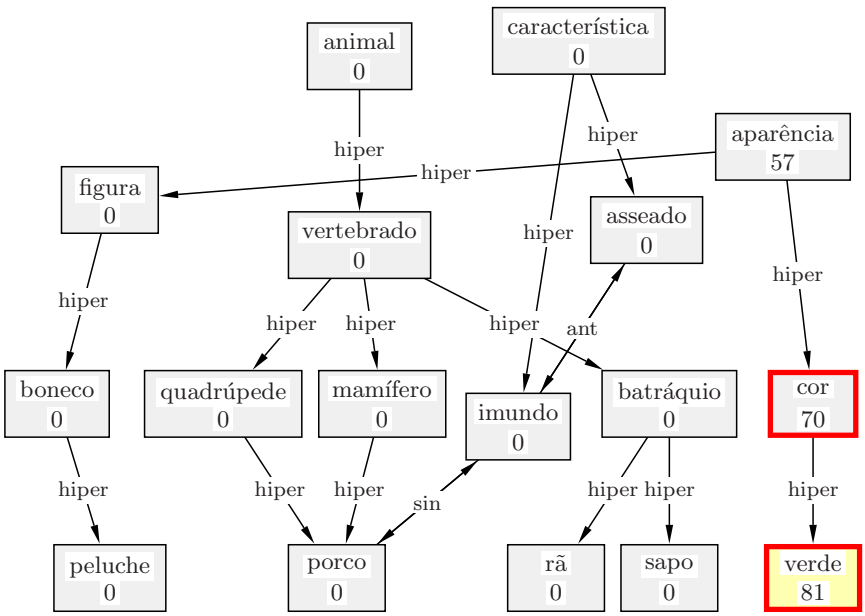

Figura 3.20: Espectro de activação semântica (resposta *verde*)

3.2.3 Reforço de confiança por semelhança de espectro

Ao comparar espectros de activação semântica identificam-se os nós activos que são comuns a várias respostas. Quanto mais nós activos existirem em comum maior será a afinidade semântica entre as respostas.
Voltando ao exemplo, encontramos um nó activo comum aos espectros das respostas *rã* e *sapo*, respectivamente nas figuras 3.18 e 3.19. Para o espaço de resultados em causa, estas respostas têm um espectro semântico semelhante. Esta correlação entre as respostas é vista como um reforço mútuo e ambas são valorizadas com um incremento da respectiva pontuação.

3.2.4 Espectro combinado e reforço de confiança

Para além da análise de espectros de activação por resposta, é também possível aplicar a técnica de propagação de activação sobre todos os nós resposta para obter um espectro de activação semântica combinado. Daqui resulta um grande número de nós em estado activo, contudo o relevante neste procedimento é seleccionar os nós que não correspondem a respostas e têm o maior nível de activação. Estes nós serão os que mais beneficiaram de propagação, eventualmente por estarem associados a muitos nós resposta, ou por estarem associados a respostas que à partida tinham elevada pontuação. Em qualquer dos casos, esses nós de maior activação representam uma tendência ou um consenso entre os significados de todas as respostas consideradas. Por conseguinte, as respostas associadas a nós que no espectro combinado se situem junto aos nós de maior activação devem beneficiar de um incremento na pontuação. Este incremento pode ser gradual em função da distância ao nó mais activo.
A figura 3.21 representa o espectro de activação combinado para o espaço de resultados do exemplo.
Entre os nós activos destaca-se *batráquio*, com nível de activação final 78. As respostas que lhe estão associadas são *rã* e *sapo*, que assim recebem uma bonificação na respectiva pontuação (pela segunda vez) e emergem como respostas preferidas pelo sistema.

3.2.5 Identificação de novas respostas no espaço de resultados

Consideremos os cenários em que a pergunta em causa é do género "*Que tipo de Conceito é Entidade?*" e o espaço de resultados inclui uma resposta R e ainda um nó com *Conceito*, seu hiperónimo. Se houver outro nó N encadeado, tal que N é hiperónimo de R e *Conceito* é hiperónimo de N, então N constitui uma resposta possível. Transpondo esta regra para o exemplo, se considerarmos válida a resposta *sapo*, então podemos admitir também como respostas os seus hiperónimos abaixo de *animal* (termo referido explicitamente na pergunta). No caso, essas respostas seriam *batráquio* e *vertebrado*.
O grau de confiança associado a estas novas respostas candidatas é influenciado pelas respostas iniciais, em cujo espectro são nós activos. O espectro combinado denota essa

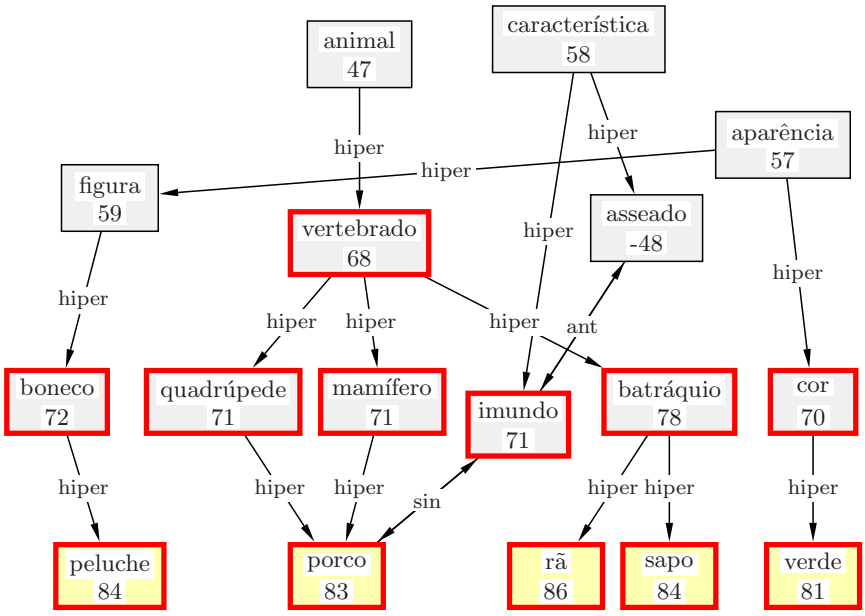

Figura 3.21: Espectro resultante da activação semântica de todas as respostas

influência das respostas candidatas base nos nós intermédios através do nível de activação, que é assim usado como pontuação para estas novas respostas.

3.3 Funcionamento do Sistema de Pergunta-Resposta

De modo simplificado, o objectivo do sistema é receber uma pergunta sob a forma de texto, em Português, e retornar uma resposta exacta. Na procura da resposta são encontrados outros resultados, alguns dos quais podem eventualmente ser tão ou mais acertados que a resposta seleccionada. Neste modelo prevê-se que o sistema apresente ao utilizador também as melhores respostas alternativas, se existirem, numa lista de resultados organizada por ordem decrescente do respectivo grau de confiança.
Cada resposta tem associada a frase ou expressão de onde foi captada, juntamente com uma pontuação que representa o grau de confiança que o sistema lhe atribui. Para além do excerto de texto a que a resposta está associada, há também um identificador do documento de origem, que pode ser um documento na *Web*, um texto do repositório local ou um tópico proveniente de outro sistema consultado.
Para chegar ao resultado pretendido para uma questão, o sistema realiza um conjunto de operações que compreende as seguintes fases:

1. análise da questão
2. recuperação de informação

3. extracção de respostas

4. filtragem e graduação dos resultados

5. apreciação global, promoção de respostas e convergência

O procedimento específico para cada uma das etapas é descrito ao longo das secções seguintes.

3.3.1 Análise da questão

Antes de iniciar a procura de respostas é necessário saber a natureza daquilo que se procura. Para isso, o sistema classifica a pergunta numa das categorias apresentadas na secção 3.1.5. Esta classificação requer uma análise morfo-sintáctica ao texto da pergunta, designadamente para a identificação do interrogativo e verbo principal, se existirem. O analisador Palavras[Bic00] produz uma árvore sintáctica com informação da morfologia de cada palavra.
É importante detectar o foco da questão, isto é, a entidade ou objecto de que trata a pergunta, acerca da qual se pretende determinada informação. Para facilitar a identificação de nomes ou designações compostas por várias palavras, o sistema usa uma ferramenta de reconhecimento de entidades mencionadas com dois modos de operação. Por um lado, efectua o reconhecimento de entidades através de um catálogo estático, com centenas de milhares de entradas classificadas, gerado a partir do recurso Repentino[SPC06]. Depois, para ser mais dinâmica e compensar o limite do catálogo local, esta ferramenta também valida a designação de entidades através pesquisas na *Web*, com pesquisas sobre determinados eventos como nascimento, morte ou inauguração da presumível entidade.
Nesta fase são ainda usados outros dois recursos: um dicionário de sinónimos e uma rede semântica local. Com o dicionário, os conceitos empregues nas regras de classificação da pergunta ficam automaticamente alargados. A rede semântica permite testar a compatibilidade entre conceitos, por exemplo através da relação hiperónimo.
Em suma, a análise da questão determina a categoria (ou categorias) em que a questão se enquadra. Como atributos da categoria, estarão identificados o foco da questão, entidades mencionadas no texto da pergunta, o verbo principal e um eventual complemento (possível alvo de uma acção por parte do foco), se existir. Quando aplicável, será também adicionada informação sobre a pergunta anterior do utilizador, para ajudar a contextualizar e eventualmente fornecer critérios de escolha do que será mais relevante para o utilizador, para uma fase posterior do processo.

3.3.2 Recuperação de informação

Os recursos disponíveis como fonte de respostas possuem um enorme volume de dados. O processamento minucioso da extracção de respostas seria impraticável se aplicado directamente sobre a totalidade desses recursos, especialmente nos casos de informação

não estruturada. Assim, e como sucede na generalidade dos SPR, antes de tentar identificar possíveis respostas, o sistema colecciona um conjunto de documentos que possam contribuir para encontrar uma resposta, por terem alguma relação com os elementos recolhidos antes, na análise ao texto da pergunta. Os documentos pretendidos, designados de **documentos candidatos**, resultam de um processo de recuperação de informação sobre a *Web* e sobre um repositório de textos local.

Procura na Web

A procura de documentos candidatos na *Web* faz-se através do motor de pesquisa Google. Este serviço aceita uma expressão de pesquisa e responde com uma lista de endereços de documentos *Web*. Os documentos mais interessantes nem sempre são os que resultam da pesquisa pelos termos exactos do texto da pergunta. A expressão a enviar ao motor de pesquisa deve ser composta por alguns termos na pergunta e por outros termos derivados destes ou inerentes à categoria associada à questão [ALG01]. Cada categoria de pergunta deste modelo tem um método próprio para construir a expressão a pesquisar na *Web*. A expansão dos termos de pesquisa baseia-se na consulta de sinónimos, hiperónimos e classes semânticas de que cada termo da questão é instância. Eventualmente, em vez de uma, cada categoria pode efectuar várias pesquisas no Google, para não tornar a expressão de pesquisa demasiado complexa com disjunções.
Os documentos *Web* no resultado do motor de pesquisa são descarregados e armazenados localmente, na forma de texto, e juntamente com os detectados nas colecções locais serão analisados para segmentação.

Procura em repositório local

O sistema também procura documentos potencialmente relevantes numa colecção de textos locais, como descrito na secção 3.1.4. Ao contrário do que acontece com os documentos na *Web*, para os textos no repositório local é possível controlar a metodologia de indexação e pesquisa.
O módulo responsável pela indexação é baseado no motor de pesquisa Lucene[8]. A recuperação de informação neste sistema usa uma combinação dos Modelos Vectorial e Booleano, na selecção e pontuação de documentos. Ao processo de indexação do Lucene foi adicionada uma componente de tratamento linguístico para o Português. Cada termo do texto a indexar é associado à sua raiz morfológica. Isto permite que a pesquisa por qualquer variante morfológica do termo obtenha textos em que o termo surge na forma raiz ou numa outra variante[9]. Na indexação é consultado também um dicionário de sinónimos, no sentido de associar um termo aos seus equivalentes. Assim, pesquisar por *inolvidável* terá como resultados textos que podem não incluir a palavra mas ter um seu sinónimo, como *inesquecível*. Todos os textos no repositório local são previamente

[8] Apache Lucene: http://lucene.apache.org/

[9] As variantes mais comuns para um termo são a sua forma no singular e no plural, e para ambas, o masculino e o feminino. No caso dos verbos há o infinitivo e depois todas as conjugações, nas diferentes pessoas do singular e do plural.

indexados, para serem considerados na procura de documentos candidatos a contribuirem com respostas.
A expressão de pesquisa a passar ao Lucene é gerada por cada categoria de questão, como acontece para a pesquisa na *Web*. Mas neste caso não é necessário incluir os sinónimos na expansão dos termos, pois isso já foi contemplado na indexação. Após a primeira tentativa de encontrar os documentos candidatos, se não houver resultados da pesquisa do Lucene procede-se a uma pesquisa menos exigente, que inclui o texto da pergunta.

Segmentação e selecção de trechos

Alguns documentos candidatos podem ter dezenas de páginas de conteúdo, outros menos, mas em geral muito mais texto que o necessário para retirar uma expressão aceitável como resposta. A segmentação e selecção de trechos[10] é um processo de procura de excertos relevantes dentro do próprio documento candidato. É sobre estes trechos que incide o processamento intensivo da fase de extracção. Deste modo, evita-se a aplicação de operações computacionalmente dispendiosas nas partes em que não é detectada nenhuma relação com a pergunta [LVF02]. Cada documento é dividido numa lista de segmentos, que correspondem a frases do texto. Nessa lista assinalam-se os segmentos que incluam termos usados na expressão de pesquisa, do passo anterior, que são os termos da pergunta ou derivados destes, semanticamente compatíveis.
Muitas vezes a resposta não está num dos segmentos assinalado, mas noutro segmento próximo. Para cada segmento assinalado, forma-se um trecho de texto que vai incluir:

- o segmento anterior
- o segmento actual
- os dois segmentos seguintes

Daqui resulta que pode haver sobreposição entre a janela textual de cada trecho. Assim acontece se um dos dois segmentos seguintes for também seleccionado por incluir termos relacionados com a questão.
Há casos em que o segmento assinalado é uma frase com uma anáfora, relativa a alguma entidade mencionada no texto precedente. É o caso de listas em que se enumeram as actividades desenvolvidas por uma instituição ou por uma pessoa célebre, algo frequente em documentos da *Web*. Para estes casos o trecho passa a incluir outros segmentos precedentes ao assinalado, até um deles incluir no seu texto uma entidade. Por vezes o primeiro segmento pode ser o título de uma secção.
Algumas categorias de pergunta dão origem a trechos adicionais, resultantes de pesquisas em recursos específicos. Dicionários e enciclopédias *online* podem fornecer uma possível definição para um conceito. Para as perguntas acerca de locais também pode obter-se informação geográfica em fontes especializadas.
Cada trecho extraído é representado por um objecto que, para além do texto seleccionado, inclui também um identificador do documento de origem, bem como o título

[10] Processo também designado por *passage retrieval*.

daquele, se estiver identificado pelo sistema, e ainda a posição em que o mesmo se encontrava no resultado da recuperação de documentos. Estes elementos são considerados adiante, na fase extracção de respostas, para cálculo da pontuação base a atribuir a determinados resultados.

3.3.3 Extracção de respostas

Em geral, a resposta candidata é uma palavra ou uma expressão de algum dos trechos. Para ser considerada, deve ter o tipo esperado para o resultado (uma quantidade, um local, uma categoria semântica) e surgir associada ao foco da questão, através de um verbo ou característica mencionada directa ou indirectamente na pergunta.
As regras concretas para a procura de respostas num trecho variam em função da categoria de pergunta. As asserções no texto que relacionam a resposta com a pergunta têm, muitas das vezes, uma estrutura frásica característica do tipo de resposta previsto para a categoria. A extracção de resultados, ainda provisórios, pode basear-se na ocorrência de padrões relativos ao foco da pergunta e a estruturas sintácticas prováveis para as respostas [Sou01] [HHR02].
Cada trecho seleccionado na fase anterior é agora analisado por regras de extracção específicas da categoria da pergunta. Cada regra é especializada em verificar um conjunto de requisitos para localizar uma resposta candidata no texto. Para a categoria *Definição*, por exemplo, a primeira regra procura uma forma muito simples para definir conceitos: *Conceito Ser Resposta*. Se pretendermos a definição de *"Fogo de São Telmo"*, então esta regra extrai uma resposta de frases como:

> *O fogo de São Telmo é um tipo de fogo fátuo que consiste numa descarga electroluminescente.*

Esta categoria recorre também à pesquisa no dicionário, para além de oito regras de procura superficial de padrões de resposta no texto, aplicáveis sobre documentos de qualquer fonte. Algumas regras são relativas a disposições textuais que podem incluir mas não implicam, necessariamente, a presença de uma resposta. Isto significa que o resultado provisório pode não ser válido, pelo seu significado ou pela forma da expressão correspondente. Aqui, a regra atribui ao resultado uma pontuação inferior à que resulta dos padrões mais plausíveis.
Algumas regras inferem respostas a partir de dados encontrados nos textos. Para a categoria *QuantidadeIdade*, os valores detectados para a idade pretendida têm de ser colocados em perspectiva, considerando o tempo decorrido até ao presente. Quando a data de referência é relativa ao nascimento, por exemplo, então a resposta é determinada pelo número de anos decorridos desde então. Se esse valor for zero então calculam-se os meses, ou eventualmente os dias. Estes cálculos no eixo do tempo são realizados em função dos momentos chave para a percepção temporal, descritos aquando do contexto temporal, na secção 3.1.2 (página 33).
A categoria *Quando* também usa regras para obter indirectamente respostas em função de referências no texto. Para determinar quando ocorreu um evento, para além das regras que procuram por datas concretas, há ainda regras para gerar resultados em função

de indicações temporais relativas. Para tal, é importante ter o contexto temporal do texto analisado. O sistema leva em consideração expressões como *"no ano passado"* ou *"este mês"*, entre outras. Inclusivamente, o tempo verbal usado junto à expressão temporal relativa vai ditar o sentido em que deve ser feito o cálculo ou ajuste, para o passado ou para o futuro da data de referência. No exemplo: *"Quem é o actual campeão olímpico de triplo salto?"*, o termo *actual* é associado ao momento presente. Existirão muitos documentos com a expressão *"actual campeão"*, mas são relativos a várias épocas.

Consideremos outro exemplo, para a categoria *Confirmação*. Quando a pergunta é relativa à acção praticada por uma entidade sobre outra entidade, há uma regra que procura essas entidades e o verbo que descreve a acção, ou outro seu sinónimo. Esta regra tenta confirmar o exposto na questão. Para o mesmo caso, outra regra procura o mesmo género de frase, mas com uma expressão de negação associada verbo, ou onde o verbo presente é antónimo do verbo que descreve a acção na pergunta. Esta última regra dá lugar às respostas "*não*".

Há respostas que não se detectam através do simples processamento de uma frase. Por vezes é preciso tirar ilações, conjugando várias evidências. Por exemplo, no processamento da pergunta *"Em que continente nasceu Vanessa Fernandes?"*, o sistema não encontrou na análise superficial do texto nenhuma asserção explícita acerca da atleta ter nascido na Europa. A única indicação sobre o seu nascimento foi: *"Vanessa Fernandes nasceu em Perosinho a 14 de Setembro de 1985"*. Neste caso é necessário procurar informação sobre Perosinho, chegando a Portugal e concluindo finalmente que o local de nascimento se situa na Europa, que é um continente e, portanto, constitui uma resposta. Esta inferência é baseada em regras que relacionam elementos das pergunta, e também de textos, com factos da base de conhecimento inicial, informações de ontologias e outras fontes de informação estruturada. A automatização deste processo é delicada e a exploração das diferentes "pistas" prejudica o tempo de resposta. Para impedir que este processo tente seguir uma cadeia de indícios demasiado longa e demorada, eventualmente infinita, as regras de inferência limitam o número de tentativas e o número de elementos considerados em cada tentativa. Um elemento considerado numa tentativa pode ser, por exemplo, uma relação semântica entre uma entidade e um conceito, expressa numa rede semântica.

Nas categorias em que as respostas podem ser frases ou expressões de comprimento indeterminado, é necessário decidir onde termina a resposta a extrair, em particular para os casos em que a frase é demasiado extensa. Pela análise sintáctica do texto, escolhe-se o ponto de divisão entre grupos sintácticos após terem sido recolhidos elementos suficientes para a resposta. Esses elementos suficientes variam também em função da categoria. Como exemplo, para as definições de conceitos não verbais é necessário pelo menos uma palavra do tipo *substantivo*.
Uma resposta pode ser encontrada por regras diferentes, possivelmente em vários trechos. Cada resultado é associado à sua lista de trechos de suporte, que por sua vez têm

a identificação dos documentos de origem.

O procedimento adoptado na fase de extracção visa recolher o maior número possível de eventuais respostas. A flexibilidade que aqui é concedida dá lugar a muitos resultados incorrectos. Portanto, é necessário filtrar os casos que claramente não constituem uma resposta válida, como se descreve na secção seguinte.

3.3.4 Filtragem e graduação dos resultados

Há diversos motivos para um resultado não ser aceitável. Este primeiro nível de validação vai descartar as expressões cuja constituição seja inadequada para uma resposta, ou aquelas em que a categoria semântica é manifestamente incompatível com a categoria de pergunta.
Devido à grande variabilidade de construção frásica, inerente à língua natural e acentuada também pela heterogeneidade nas fontes dos textos, o valor de resposta capturado pelas regras da fase de extracção pode apresentar uma forma diferente do esperado. Efectuando uma verificação simples, eliminam-se casos em que a expressão inicia com sinais de pontuação ou outros símbolos estranhos à categoria de resposta. É também aqui que se filtram respostas vazias ou demasiado compridas para a categoria.
Os resultados provisórios são ainda objecto de outra verificação, desta vez relativa à natureza do que representam. Mesmo sem uma análise semântica aprofundada, há alguns testes de compatibilidade com a categoria de pergunta que evidenciam casos a excluir do conjunto de respostas. Se um resultado contém um nome onde se espera uma definição, se existe uma quantidade expressa em dígitos proposta como solução para uma pergunta do tipo *Quem*, são algumas situações concretas de resultados a descartar.

Para além da lista dos trechos que contribuíram para cada resposta, importa também quantificar o grau de confiança estimado para cada uma delas. Para isso, atribui-se uma pontuação a cada resultado. Algumas metodologias baseiam-se na frequência da resposta no conjunto de todos os trechos, ou eventualmente na distância (mínima ou média, em palavras) até ao elemento central do trecho [CCL01], ou ao conceito ou entidade que é o foco da questão.
Neste modelo, é a categoria de pergunta que determina a pontuação base para as respostas que as suas regras de extracção originam: PbC. Cada regra de extracção é especializada num procedimento ou numa particular configuração sintáctica, que pode ser mais ou menos segura para encontrar resultados para a categoria de questão. Em função disso, há uma parte da pontuação que advém da regra, proporcional à taxa de acerto estimada para o procedimento que a mesma efectua: PR. Na aplicação de uma regra de extracção a um trecho, a observação de nomes, termos, tempos verbais em concordância perfeita com o texto da pergunta pode originar uma bonificação PB. Por outro lado, outros aspectos de desvio relativamente à concordância perfeita podem originar alguns pontos de penalização: PP. É o que acontece se a resposta é extraída de um bloco de texto que contém apenas alguns elementos chave da pergunta, ou quando há texto

supérfluo entre o foco da questão, ou o verbo, e a zona da resposta. Também pode haver dedução de pontos se o foco da questão é uma entidade com uma designação formada por várias palavras mas apenas a primeira e a última estão presentes, algo frequente nas referências a pessoas pelo seu nome.
A pontuação inicial associada a uma resposta candidata é:

$$Pontos = PbC + PR + PB - PP \tag{3.3}$$

Em caso de empate na pontuação das melhores respostas, ainda se pode fazer a diferenciação pelo número de trechos de suporte de cada resultado.
Nesta fase do processamento de uma pergunta já existe uma lista ordenada de resultados. O passo seguinte é uma revisão dessa lista com o objectivo de melhorar a ordenação das respostas, para uma sequência globalmente mais correcta.

3.3.5 Apreciação global, promoção de respostas e convergência

Na última fase do tratamento de uma pergunta, o sistema tenta imitar o comportamento inteligente de um humano quando perante várias opções toma uma decisão fundamentada. Seguindo os princípios enunciados na secção 3.2, reaprecia-se cada resposta candidata, analisando o seu valor e respectivos contextos espacial, temporal e semântico, com critérios de preferência que incrementam a pontuação de alguns resultados.

respostas para o tipo *QueQual*

O primeiro passo é construir o espaço multidimensional de respostas para a pergunta, representando lá os resultados provisórios advindos das fases anteriores. Com o contexto das respostas representado em cada eixo do espaço geral, aplicam-se critérios de escolha ao longo de cada uma das dimensões: temporal, espacial e semântica.
Têm prioridade as respostas cujo tempo de asserção ou período de validade seja compatível com ou próximo ao tempo da resposta. Quando não existe uma imposição temporal escolhe-se a resposta mais próxima do presente. Do mesmo modo, se a pergunta é sobre um facto ou característica para determinada região geográfica, então a atenção deve recair nos resultados cujo contexto espacial é ou está inserido nessa região.
Importa esclarecer que há resultados cujo contexto temporal ou espacial é indeterminado. Isto acontece quando o sistema não encontra evidências dessas dimensões, tanto no texto como em metainformação associada aos documentos. Como não há forma de relacionar estes contextos com o previsto para a pergunta, as respostas nestas condições são sempre consideradas, isto é, não recebem pontos mas também não são descartadas nesta parte do processo.

Na categoria *QueQual* as respostas são curtas, muitas vezes formadas por uma só palavra, e enquadram-se em determinado género conhecido. O contexto semântico de cada factóide é representado pela teia de associações que derivam do nó que os representa no espaço de resultados, como ilustrado na figura 3.10. A afinidade semântica entre um grupo de respostas reforça a pontuação desse grupo. Esta eventual semelhança de significados é analisada com a técnica de propagação do nível activação de cada nó de resposta pela rede semântica, explicada na secção 3.2.2. São promovidas as respostas que tiverem um espectro de activação com mais nós activos em comum com espectros de outras respostas. Outra técnica, independente da anterior mas igualmente baseada no algoritmo de propagação, incrementa a pontuação dos resultados associados ao nó mais activo no espectro de activação combinado.

Para além dos contextos, há ainda a técnica de análise do valor isolado de cada resposta. Esta técnica complementar aplica um procedimento independente, sendo eficaz mesmo quando os factóides são relativos a nomes, situação em que a análise semântica não ajuda.
Para a confirmação do valor da resposta noutras fontes, realizam-se pesquisas que juntam o factóide X à descrição na pergunta. Se procuramos a montanha mais alta do México, podemos testar o resultado X com pesquisas como *"X é a montanha mais alta do México"* e outras combinações com estes elementos.

O efeito da aplicação de cada técnica em particular é apresentado no capítulo 4. Será descrita uma avaliação quantitativa das respostas que o sistema escolhe para as perguntas, antes e depois de aplicar o procedimento de cada técnica.

respostas para o tipo *Definição*

As respostas candidatas para questões do tipo *Definição* têm estrutura e comprimento variáveis. Para a mesma questão podem existir resultados bastante distintos na forma, mas com significado aproximado. O processo de representação destas expressões no eixo semântico do espaço de resultados visa esbater essas diferenças na morfologia dos termos ou na estrutura sintáctica da expressão. Através da identificação dos termos de núcleo nos resultados, constrói-se a malha de nós com a semântica que o sistema atribui a cada resposta, tal como ilustra a figura 3.12.
Embora os nós de resposta sejam relativos a expressões mais compridas, foram associados aos restantes nós do espaço com a mesma metodologia e através do mesmo conjunto de relações semânticas que os factóides. Portanto, pode usar-se a propagação de nível de activação e promover respostas com as duas técnicas: por comparação de espectros de resposta e por proximidade ao nó mais activo do espectro combinado.

Ainda no âmbito da afinidade entre resultados, foram experimentadas outras técnicas de reconhecimento de semelhanças entre respostas para reforço mútuo de confiança. Estas experiências assentam na frequência dos termos de uma resposta no texto de todas as

respostas candidatas, com o intuito de dar mais pontos aquelas que incluem as palavras mais populares no universo das respostas. A versão base incrementa a pontuação de uma resposta por um valor proporcional à frequência média dos termos que a compõem. Noutra variante desta técnica, retiram-se as *stop words* e palavras repetidas no texto da própria resposta, mantendo-se a mesma fórmula para o incremento. A variante seguinte nesta sequência acrescenta uma alteração. Para as palavras a considerar, a frequência a verificar é a do termo base da palavra (a sua raiz morfológica) e a pontuação a atribuir é agora indexada ao somatório de frequências dos termos normalizados da resposta sobre o número total de termos normalizados, pertencentes a qualquer resposta.

A confirmação por pesquisa na *Web* foi também testada para esta categoria de perguntas, adaptando o procedimento às características das respostas, designadamente o facto de possuírem uma extensão variável. A primeira variante consiste na atribuição de um valor fixo de pontos se houver um ou mais resultados na pesquisa pelo foco da questão (o conceito a definir) juntamente com a possível resposta, entre aspas. Noutra variante, a parte da resposta candidata usada na expressão a pesquisar é formada pelo seu prefixo, com tamanho aproximado para todos os resultados em análise. Para cada resposta, o texto que a representa é formado pelas primeiras N palavras não *stop words*.

Outra das técnicas de apreciação individual de respostas incide sobre a forma de cada resultado. Algumas respostas para a categoria em causa são extraídas após o verbo ser, por exemplo. Acontece que nem sempre o texto obtido corresponde ao esperado. Na definição de conceitos é usual a presença de substantivos, em particular no início[11] da definição. A primeira variante desta técnica atribui pontos às respostas constituídas por, pelo menos, N substantivos. A variante seguinte tenta minimizar o impacto da dimensão do texto e determina os pontos em função de uma taxa mínima de nomes por número de palavras na resposta. Uma terceira variante dá pontos apenas se existir um número mínimo de substantivos logo no início da possível definição.

respostas para o tipo *Quem*

Esta categoria tem a ver com associação entre entidades e determinados critérios. Na maior parte dos casos as respostas têm a identificação de uma entidade que obedece ao critério da pergunta, possivelmente o nome de uma pessoa ou de uma instituição. As respostas com nomes não têm matéria para análise semântica. A apreciação destes resultados resume-se a dois aspectos:

- a concordância espácio-temporal do contexto de cada resposta com a pergunta
- a confirmação da designação como resposta, através de pesquisas formuladas de acordo com o tipo de questão e que combinam essa designação de entidade com o critério pedido

[11] Excepto na definição de conceitos verbais, que muitas vezes inicia com um verbo. Mas é comum ter definições a iniciar com um substantivo, ou um artigo ou adjectivo seguido de substantivo.

Quando a pergunta parte do nome, o que sucede na categoria *QuemSerNome*, a resposta é uma descrição cuja estrutura é semelhante a uma definição. Aqui já se aplicam os procedimentos de análise semântica e técnicas de preferência pela forma da resposta, como acontece na categoria *Definição*.

respostas para o tipo *Quê*

As respostas para questões do tipo *Quê* podem ser designações ou descrições curtas alusivas a objectos inanimados ou entidades não humanas associadas a acções ou factos. Para além da análise dos contextos espacial e temporal, transversal a todas as categorias, aplicam-se as técnicas já descritas para as descrições e a confirmação do valor por meio de pesquisa, para todas as respostas.

respostas para o tipo *Quantidades*

Os resultados propostos como resposta a uma questão deste tipo são valores numéricos, com ou sem unidades. Para além da expressão original com a quantidade, tal como foi captada nas fontes, cada resposta inclui a avaliação da respectiva expressão, expressa num formato normalizado[12].
Em termos abstractos, uma quantidade é analisada como se fosse um factóide textual com um conceito. A diferença está na análise do significado, que aqui é mais simples e não envolve a propagação de activação. O eixo semântico em que as respostas são representadas, tal como mostra a figura 3.6, é efectivamente unidimensional e contínuo. Assim, a atribuição de pontos a respostas com afinidade no significado é exequível também nesta categoria. Podemos definir a proximidade semântica como a proximidade entre os valores das respostas. Mas se houver unidades associadas só podemos comparar grupos de respostas com a mesma unidade no formato normalizado (e que já pode ter resultado de uma conversão automática).
A técnica base consiste em bonificar uma resposta em função da existência de outros resultados na sua vizinhança V. Uma variante desta técnica restringe a atribuição de pontos às respostas que na sua vizinhança V possuem resultados com pontuação base acima da média. A intenção é destacar as respostas influenciadas pelos resultados provisórios com maior grau de confiança.
Para além de respeitar diferenças inerentes à natureza da quantidade (peso, preço, comprimento e outras), a escala empregue neste procedimento tem de ser pensada para cada pergunta, em função da dispersão das suas respostas candidatas. Se assim não for, há o risco de se obter uma vizinhança V que pode abranger todas as respostas encontradas, ou nunca incluir nenhuma, o que é igualmente inconveniente. Deste modo, a vizinhança a considerar tem um valor variável, determinado por dois factores:

[12]Durante a fase de extracção, converte-se a quantidade expressa com dígitos ou por extenso para o formato numérico base do sistema, normalizando o separador decimal e dos milhares, para assegurar uma leitura correcta aquando da avaliação da resposta.

- o valor particular da resposta em análise - uma vizinhança com $V = 2$ relativa a um valor 100 é aceitável, mas já não parece razoável se aplicada a uma resposta com valor $0,7$. Por isso determina-se V em função de uma percentagem do valor de cada resposta. Esta percentagem é definida na categoria específica de quantidade associada à pergunta.

- desvio padrão - usa-se o desvio padrão como valor máximo para V. Consideremos uma distribuição de resultados com desvio padrão 10 onde usamos a taxa 3%. Para uma resposta com valor 50 temos $V = 1,5$; para outra resposta com valor 900 temos $V = 10$. O segundo caso reflecte o uso do desvio padrão como limite superior, pois 3% de 900 é 27.

respostas para o tipo *Quando*

Uma pergunta com a categoria *Quando* tem um conjunto de respostas candidatas constituído por datas, mais ou menos precisas, ou expressões alusivas a épocas ou momentos. A técnica aplicada nesta categoria é focada nas datas por serem o tipo de resposta mais comum, ainda que, como se relata adiante, haja a possibilidade de análise semântica de expressões como *"à hora de jantar"*.
As datas encontradas nem sempre têm o mesmo grau de exactidão. Uns resultados indicam um ano, outros são um pouco mais concretos e apontam para um mês de determinado ano, e as respostas mais precisas referem-se a datas completas, com dia, mês e ano. Para datas mais precisas contidas em datas menos precisas, basta comparar a parte comum. Os resultados *"5 de Outubro"*, *"dia 5"* e *"5 de Outubro de 1143"* reforçam-se mutuamente. Entre duas datas como *"5 de Outubro de 1143"* e *"1143"*, é necessário ver se o ano da data completa é o mesmo da data menos exacta que só inclui o valor do ano. Se as datas não estão sobrepostas é necessário avaliar a distância entre elas. A interpretação feita do valor das datas tem de contemplar a quantificação da distância que as separa. Num conjunto de respostas de formato heterogéneo, é preciso encontrar a mínima unidade de tempo em comum, por vezes *mês*, mas na maioria das vezes *dia*. É com esta unidade que se avalia a distância entre todas as respostas não sobrepostas, como *2008-11-01*, *2010* ou *Julho de 2009*. E é desta forma que as datas são dispostas ao longo do eixo semântico do espaço de resultados, como ilustra a figura 3.5.
A técnica de valorização de respostas de significado aproximado é idêntica à técnica base nas quantidades, com a particularidade de algumas destas respostas corresponderem a intervalos entre dois valores (cuja unidade é *dia* ou *mês*).

respostas para o tipo *Localização*

A verificação dos resultados para esta categoria compreende duas vertentes. Há a vertente puramente semântica, ao validar a resposta como uma localização, e existe também uma análise da proximidade ou associação geográfica entre as respostas. A primeira

resume-se ao teste de compatibilidade com um dos conceitos *local*, *área geográfica* ou *posicionamento*. Mas esse procedimento já foi concretizado na fase de extracção, juntamente com a verificação de algum critério eventualmente exigido pela questão. A segunda vertente, baseada na afinidade posicional ou geográfica entre resultados, é usada nesta fase de bonificação de respostas. Através da base de conhecimento inicial do sistema e da consulta de sistemas de informação geográfica, constrói-se uma malha de resultados, como acontece para os factóides em geral. Desta vez, a malha representativa da dimensão semântica do espaço de resultados de uma pergunta será composta por relações *instânciaDe* (exemplo: instâncias de *cidade*, *país*) e *localizadoEm* (expressando a região mais abrangente em que cada local se insere). Sobre esta malha de resultados, aplicam-se as técnicas de propagação do nível de activação desde os nós de resposta. Por exemplo, duas respostas candidatas, uma com a povoação A e outra com a província ou estado a que pertence a primeira, B, devem assim ser bonificadas se os restantes resultados são povoações dispersas e pertencentes a zonas diferentes.
Outra técnica aplicável para diferenciação entre respostas baseia-se na confirmação por pesquisa na *Web*. Para além disso, se existir informação que caracterize o contexto temporal das respostas, a preferência recai sobre o resultado mais actual.

Por haver alterações sucessivas na pontuação de algumas respostas e, consequentemente, na organização da lista de resultados, por ordem decrescente da mesma pontuação, diz-se que o sistema está a convergir gradualmente para um resultado.
Efectuadas as várias operações de escolha e promoção de respostas, algumas delas passam a estar em evidência por terem emergido entre as demais com as bonificações que receberam. Por fim, o processo termina com a apresentação da resposta mais cotada, juntamente com a lista das melhores respostas alternativas.
A próxima secção recapitula e descreve esquematicamente os módulos mais relevantes do sistema.

3.4 Arquitectura

Explicada que está a metodologia do sistema, é altura de descrever a arquitectura com que foi concebido. Primeiro será apresentada a lista de módulos chave para o serviço de pergunta-resposta. Em seguida indicam-se alguns aspectos técnicos de implementação desse serviço e da respectiva interface para o utilizador.

3.4.1 Módulos do sistema

As fases principais no processamento de uma pergunta, apresentadas na secção 3.3, podem ser ajustadas de modo independente. Daí resulta que o sistema deve ser suficientemente modular, para que a actualização de um procedimento particular possa efectuar-se sem maiores consequências para os procedimentos seguintes. Assim, a funci-

onalidade de cada uma das referidas fases é assegurada por um módulo dedicado, como ilustra a figura 3.22. Os componentes de processamento da pergunta estão ao centro. Note-se como todos tiram partido de dados sobre as categorias de pergunta previstas no sistema. As caixas com cor azul representam as fontes de resposta. Algumas destas fontes são consultadas em múltiplas ocasiões por tarefas associadas ao tratamento da mesma questão, para a validação ou a ponderação de respostas.

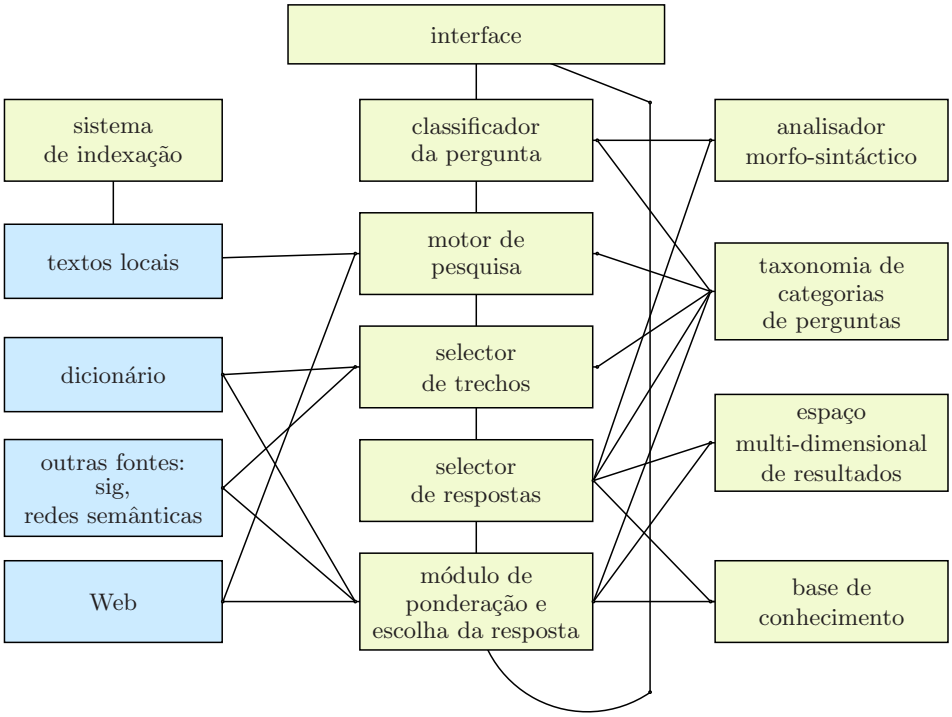

Figura 3.22: Componentes do sistema de pergunta-resposta

3.4.2 Infraestrutura para o serviço

Para implementar um serviço deste género de modo a assegurar um funcionamento eficiente, há que considerar os aspectos:

- ferramentas necessárias: que recursos ou ferramentas são consultados localmente?
- comunicação: com que sistemas é estabelecida uma comunicação?
- tempo de resposta: a recuperação de documentos, acessos à rede e análise morfo-sintáctica tendem a atrasar a resposta do sistema
- paralelismo: permitir o processamento simultâneo de duas questões, ou acelerar a resposta a uma só pergunta executando tarefas em paralelo, quando possível

- interface com o utilizador: convém separar a apresentação da funcionalidade

A solução escolhida no âmbito deste trabalho é uma arquitectura distribuída, baseada na tecnologia de *middleware* `Java RMI`[13]. Com esta tecnologia temos a expressividade da orientação por objectos a nível distribuído, com a vantagem de nos podermos abstrair de eventuais diferenças na plataforma (hardware ou sistema operativo) das máquinas usadas. Algumas ferramentas usadas pelo sistema podem restringir esta abstracção, impondo a existência de determinado ambiente, nomeadamente o sistema operativo. O sistema foi desenvolvido e colocado em funcionamento em máquinas que operam com o sistema Alinex[14].
A figura 3.23 mostra o carácter distribuído do sistema. O módulo mais exposto do sistema fica junto a um servidor *Tomcat*[15], que é consultado por aplicações comuns de navegação na *Web*. Um ou mais servidores internos asseguram a parte funcional do serviço. Estes são contactados por invocação remota de métodos sobre os seus objectos, que por sua vez poderão usar recursos locais ou remotos através da rede.

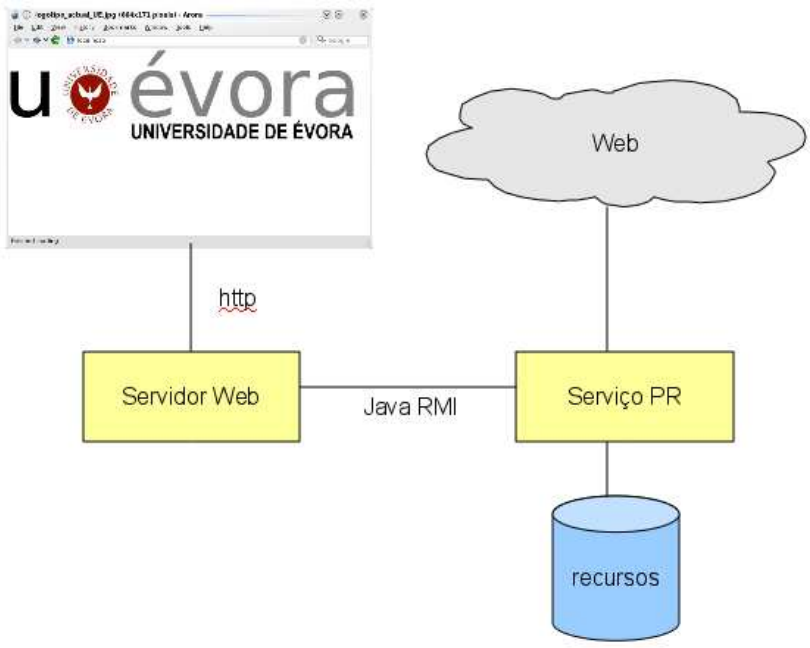

Figura 3.23: Arquitectura distribuída

[13]Java RMI: tecnologia Java para o paradigma da invocação remota de métodos (*Remote Method Invocation*)
[14]Alinex: www.alinex.org - sistema *Linux* da família *Debian* usado na Universidade de Évora
[15]http://tomcat.apache.org - servidor *Web* com funcionalidades *server side* Java

3.4.3 Interface para o utilizador

Num sistema de pergunta-resposta espera-se que a interacção com o utilizador seja simples e fácil de entender por pessoas sem grande preparação técnica. Este sistema é acedido através de uma interface *Web*, como é já comum nos dias que correm.
A figura 3.24 mostra a área em que o utilizador insere a sua pergunta. Desse mesmo ponto é possível listar e procurar questões anteriores.
O administrador do sistema pode acompanhar detalhadamente os passos intermédios

Figura 3.24: Interface: zona de escrita da pergunta

do processamento de uma questão, o que é útil para testar os diferentes módulos. Na figura 3.25 temos algumas dessas opções de consulta, nomeadamente a análise sintáctica da pergunta ou o resultado da recuperação de documentos sobre a colecção local de textos.
O resultado do processamento de uma pergunta inclui a melhor resposta do sistema e uma lista ordenada com respostas alternativas. A figura 3.26 tem a informação apresentada ao utilizador como resultado do processamento da questão

Que prémio recebeu José Saramago?

Podemos observar a resposta escolhida, *Nobel*, e com menos um ponto, a melhor alternativa: *Prémio Nobel da Literatura*. Cada resposta é acompanhada por frases ou excertos que a suportam. Do lado direito existem apontadores para o utilizador ler os documentos na íntegra, caso queira confirmar o contexto de origem da resposta. Estes apontadores podem levar a documentos da *Web* ou a documentos do repositório local.

Figura 3.25: Interface: diversas opções de consulta

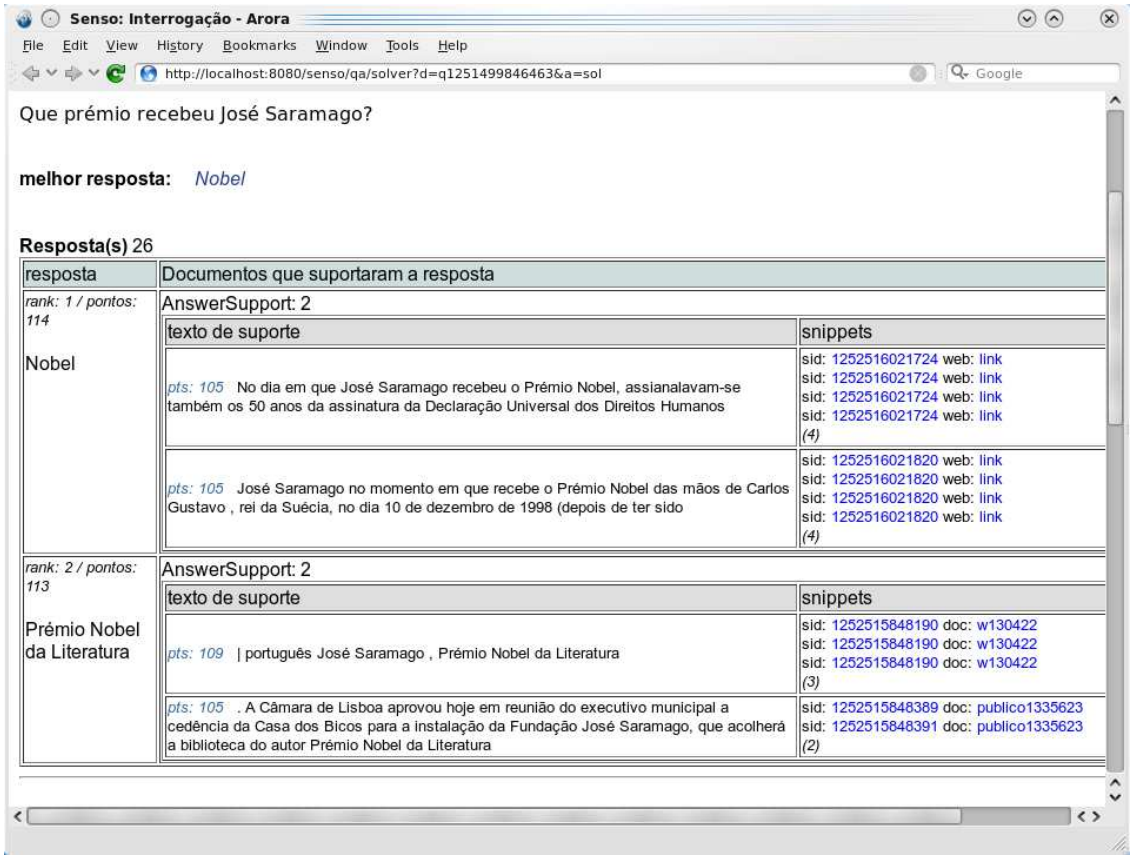

Figura 3.26: Interface: resposta para uma questão

3.5 Resumo

Neste capítulo descreve-se o modelo proposto por esta dissertação para sistemas de pergunta-resposta de âmbito geral. Tomando como inspiração a propagação de activação no processo de recuperação de informação na memória humana, o modelo desenvolvido procura contextualizar respostas candidatas no tempo, no espaço e também numa dimensão semântica, para daí obter, ponderadamente, um resultado.
Foi explicado o funcionamento de um sistema desenvolvido com esta metodologia, incluindo as técnicas usadas na selecção de respostas. A representação das respostas candidatas no espaço multidimensional permite uma análise de cada resultado tendo em conta as informações sobre todas as alternativas encontradas. A identificação de padrões num conjunto de respostas, ou a proximidade das respectivas representações ao longo de várias dimensões do espaço, pode motivar um reforço da confiança nessas respostas. Isto conduz à promoção de respostas através do incremento da respectiva pontuação, no sentido de convergência para um resultado que se espera ser mais correcto.

O próximo capítulo relata a avaliação efectuada ao sistema desenvolvido com este modelo.

Capítulo 4

Avaliação do Modelo

Durante este trabalho foi implementado um sistema de pergunta-resposta de acordo com o modelo teórico referido no capítulo anterior. Como instância concreta e que materializa as ideias do modelo geral, o sistema permitiu a aplicação prática dos procedimentos propostos, cujos resultados podem ser objecto de uma avaliação objectiva.
A primeira secção deste capítulo relata diferentes experiências realizadas com o sistema. A secção 4.2 é dedicada à avaliação das técnicas propostas para selecção das melhores respostas. Para concluir o capítulo, a secção 4.3 resume a avaliação aqui descrita.

4.1 Experiências

Um sistema de pergunta-resposta é usualmente um mecanismo complexo, constituído por diversos módulos combinados para o propósito de encontrar respostas certas para as perguntas. O sistema inerente a este trabalho apresenta igualmente uma estrutura complexa, mas organizada de forma modular. Isto permite que, para além da apreciação do resultado final, se possa testar o comportamento de componentes internos, responsáveis por fases intermédias do serviço.
Esta secção apresenta várias experiências efectuadas com o sistema, em diferentes momentos e com âmbitos distintos.

4.1.1 Laboratório: respostas artificiais

Durante a fase de implementação foi necessário testar cada módulo desenvolvido, garantindo a correcção dos diversos procedimentos ao longo do processo de responder a uma questão. Uma experiência importante, designadamente para a validação do critério de selecção de resposta pelo contexto espacial e/ou temporal, foi a injecção de respostas artificiais no sistema.
Como foi referido na secção 3.1.2, os contextos espacial, temporal e semântico influenciam o modo como uma pessoa interpreta uma pergunta e, consequentemente, o processo

de escolha da resposta. Para o sistema ser sensível ao tempo e ao espaço, com vista à escolha de respostas com os critérios de concordância de contexto descritos no início da secção 3.2, é necessário que se consiga determinar para cada resposta qual o seu enquadramento ou contexto nessas dimensões. O valor captado para estes contextos de uma resposta provém de atributos associados à fonte dessa resposta, para fontes de informação estruturada, ou resulta da análise linguística do texto, no caso de fontes de informação não estruturada.
A percepção automática dos contextos temporal e espacial das respostas falha na maioria das vezes, resultando num contexto indefinido. Ainda assim, o princípio de escolha de respostas com contextualização espácio-temporal adequada continua válido. Para testar a implementação destes critérios de selecção, aplicou-se ao sistema um gerador de respostas artificiais, no final da fase de extracção. O objectivo era inserir respostas com contextos temporal e espacial definidos, facilitando a avaliação do processo de escolha que ocorre em fase posterior à extracção.
Terminada a validação do processo em cada uma dessas duas dimensões, o gerador foi desligado. A escolha do resultado actual ou mais contemporâneo, juntamente com a preferência para resultados com âmbito espacial apropriado, são critérios empregues quando possível, em função da existência de respostas com contextos determinados.

4.1.2 Utilização do sistema por humanos

Com um âmbito mais abrangente, foram realizados testes de utilização do sistema por parte de pessoas não especializadas em pesquisa de informação. O foco destes testes não foi nenhum aspecto técnico em particular, mas antes a avaliação do serviço prestado pelo sistema de pergunta-resposta como um todo.
Através de uma interface *Web* (como ilustrado na figura 3.24), os utilizadores inseriram perguntas para testar o comportamento do sistema. Com a observação de padrões na formulação das questões e também pelas críticas aos resultados encontrados, foram realizados alguns ajustes na classificação de perguntas (secção 3.3.1) e também a alguns atributos das categorias atribuídas às questões, apresentadas na secção 3.1.5.
Foi também durante esta experiência que se decidiu incluir no resultado uma lista com as N melhores respostas, para além daquela que é a resposta escolhida pelo sistema. Os utilizadores demonstraram interesse nas alternativas propostas pelo sistema, inclusivamente nos casos em que a primeira resposta é considerada válida. Do mesmo modo, a curiosidade sobre o modo como se chegou à resposta motivou ainda a colocação de um apontador para o documento de proveniência da mesma. Isto permite aceder de imediato a documentos na *Web* captados durante o processamento da questão, a textos locais ou a outros recursos consultados por forma a confirmar ou até a complementar a resposta fornecida.
O utilizador pode ainda observar o efeito das técnicas de apreciação e reordenação da lista de respostas, descritas na secção 3.3.5 e baseadas nos princípios apresentados na secção 3.2. Juntamente com o resultado do processamento de uma questão, a interface tem opções para consulta da ordenação das respostas candidatas, antes e depois

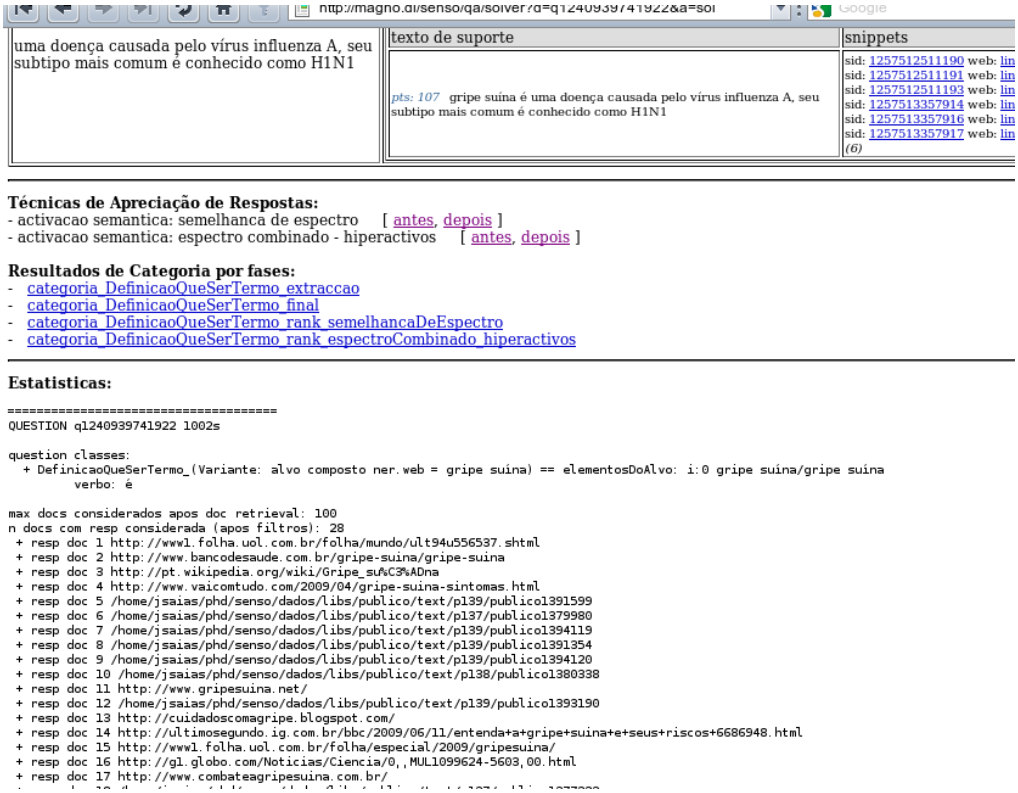

Figura 4.1: Origem de respostas e consulta de cada técnica de apreciação

da aplicação de cada uma dessas técnicas. A figura 4.1 mostra essas opções seguidas de um pequeno resumo com a classificação da pergunta *"O que é a gripe suína?"* e a identificação de algumas fontes de respostas.

4.1.3 Participação em desafios da especialidade

Num outro género de experiência, o sistema foi posto à prova numa actividade de pergunta-resposta (QA@CLEF) do *Cross Language Evaluation Forum* (CLEF). Neste desafio, os sistemas são confrontados com um conjunto de 200 perguntas de vários tipos. Não é dada a informação sobre o tipo de pergunta, que a organização classifica em *factóide*, *definição* ou *lista*, para além de perguntas para as quais não há resposta conhecida nas fontes disponíveis, as questões do tipo *NIL*. As respostas são procuradas em textos de imprensa e documentos da Wikipédia. Existem várias modalidades de participação no QA@CLEF, que consistem na combinação de idiomas usados para as perguntas e para as respostas.
Tendo havido já duas participações prévias neste desafio com representantes do Departamento de Informática da Universidade de Évora [QQR+04] [QR05], foi em 2007, após

um ano de ausência, que o sistema SENSO foi usado pela primeira vez no QA@CLEF.

QA@CLEF-2007

A participação de Évora no QA@CLEF em 2007, para o Português, foi baseada no sistema SENSO [SQ07]. Nesta versão, o sistema apresentava uma módulo de extracção de respostas com duas abordagens em paralelo. Por um lado, o processo de pesquisa textual directa, baseado na detecção de padrões sintácticos de onde se extraem respostas para determinadas perguntas. Depois, o processo lógico, para o qual era formada uma base de conhecimento dedicada à pergunta em análise, que consistia num conjunto de factos em Prolog. Os factos são uma representação semântica parcial de textos considerados relevantes para uma questão. Cada facto é uma estrutura em Lógica de Primeira Ordem (uma DRS*), gerada após a análise morfo-sintáctica do texto, com uma versão reescrita de uma frase e dos seus respectivos referentes do discurso, com base na teoria *Discourse Representation Structures* (*) [KR93]. Com o algoritmo de resolução associado ao Prolog, procuram-se soluções para a expressão lógica (DRS) da pergunta, tendo como informação a lista de factos e uma ontologia [SQ08b].
As respostas recolhidas por ambos os processos são passadas para uma só lista ordenada. A resposta do sistema para o QA@CLEF é a primeira da lista, que corresponde ao resultado com maior pontuação.
A comparação directa com resultados obtidos em anos anteriores não era relevante, devido a algumas mudanças no formato do desafio e pelo facto das perguntas serem diferentes. Mas entre os sistemas participantes na mesma tarefa de 2007, os resultados do sistema SENSO foram acima da média, com 42% das respostas consideradas correctas [GFH+08]. Foi nas perguntas do tipo *definição* que o sistema obteve melhor classificação, com 63% de resultados acertados. Já nos *factóides*, os resultados do sistema estavam correctos apenas em 38% das perguntas.
No mesmo ano, o sistema foi ainda aplicado noutra actividade associada ao QA@CLEF: o AVE[1], um exercício de validação de respostas [PVV07]. O processo de validação consistiu na comparação entre a hipótese apresentada como possível resposta e a resposta que o próprio sistema dava com base no mesmo texto.

QA@CLEF-2008

No ano seguinte houve mais uma edição[2] do desafio. A equipa de Évora participou uma vez mais na versão em Português do QA@CLEF com o sistema SENSO [SQ08a]. Em 2008 o sistema beneficiava de algumas actualizações. A recuperação de documentos foi melhorada ao nível da expansão de termos a usar para a pesquisa. Isto contribuiu para reduzir o número de falsas respostas *NIL*, de 99 para 19. Mantiveram-se as duas vertentes, lógica e de pesquisa de padrões no texto, para extracção de respostas, mas surgiu um módulo para validação das respostas extraídas. O novo componente visa reduzir

[1] Answer Validation Exercise
[2] 2008 foi o último ano da actividade QA@CLEF.

os erros, favorecendo as respostas candidatas que verifiquem critérios de popularidade através de uma pesquisa na *Web*.
Entre os sistemas das seis instituições participantes, o sistema SENSO manteve a segunda melhor percentagem de acertos sobre a totalidade das perguntas, com 46.5%, sendo ultrapassado apenas pelo sistema PRIBERAM. Tal como para a maioria dos sistemas em competição, registou-se uma pequena melhoria desde o ano anterior [FPA+08]. À semelhança do ocorrido em 2007, foi no tipo *definição* que o sistema obteve a sua melhor classificação, com 89% de respostas correctas. Para perguntas cujas respostas estavam classificadas como *factóides*, o sistema acertou apenas em 35% dos resultados.

4.2 Avaliação das técnicas de escolha de respostas

A secção 3.3.5 descreve as técnicas de apreciação de resultados usadas em cada categoria de pergunta para promover e preterir respostas candidatas, de acordo com os princípios do modelo teórico enunciados na secção 3.2. Pretende-se que estas técnicas contribuam positivamente para o desempenho do sistema, gerando uma ordenação das respostas candidatas que seja mais adequada que a anterior, decorrente da graduação efectuada na fase de extracção. Esta melhoria nos resultados ocorre sempre que uma resposta candidata correcta ultrapassar uma resposta errada na lista devolvida.

4.2.1 Mecanismo de avaliação

Para algumas categorias podem aplicar-se várias técnicas. A avaliação efectua-se sobre o resultado produzido pela aplicação cada uma das técnicas individualmente. Considera-se que a técnica melhora os resultados se os parâmetros avaliados são mais favoráveis que os registados na lista de respostas inicial.
O mecanismo de avaliação funciona da seguinte forma:

1. extracção e avaliação manual - o processo inicia com o registo de um conjunto de perguntas, que para o efeito são todas da mesma categoria. Cada pergunta nesse conjunto é processada pelo sistema (sem aplicação de nenhuma das técnicas em análise), sendo registadas numa base de dados todas as respostas e respectiva ordem no resultado da extracção. O passo seguinte nesta fase é uma avaliação humana de cada resposta, classificando-a de correcta ou incorrecta para a questão, tendo em conta a sua proveniência e os contextos. Convém que o teste seja efectuado com perguntas cuja lista de resultados inclua pelo menos uma resposta correcta, para se poder observar a evolução. Se não for assim, então a ordenação dos resultados, que nesse caso são todos errados, será indiferente. Pelo mesmo motivo, as perguntas em que a fase de extracção gera exclusivamente respostas candidatas correctas também não são úteis neste processo de avaliação.
   Esta fase termina com uma análise estatística da lista de resultados. Para cada pergunta registada verifica-se:

- a resposta do sistema, que é a primeira da lista, está correcta? (sim / não)
- índice ou posição da primeira resposta correcta (IPRC) - idealmente seria sempre 1
- índice ou posição da primeira resposta errada (IPRE) - é um indicador especialmente útil quando há várias respostas consideradas correctas. Nessa situação seria desejável maximizar este valor.
- acerto no top 5: número de respostas correctas entre as cinco primeiras da lista de resultados. O valor óptimo neste indicador é 5 em 5, sempre que a pergunta der lugar a 5 ou mais respostas candidatas, ou N em N, para $1 <= N < 5$.
- acerto no top 10: número de respostas correctas entre as dez primeiras da lista de resultados.

E de seguida obtêm-se os valores médios para o conjunto das perguntas registadas, em cada um destes cinco indicadores.

2. aplicação e avaliação da técnica - a avaliação da técnica começa com a aplicação da mesma ao conjunto de respostas candidatas para cada pergunta, já extraído e registado. A lista resultante da aplicação da técnica é novamente armazenada na base de dados. Finalmente, executa-se a mesma análise estatística sobre os novos resultados.

Os valores de cada indicador denotam a evolução entre o antes e o depois da técnica considerada. Há duas vertentes de melhoria dos resultados. Por um lado, aumentar o acerto na resposta do sistema (a resposta de índice 1 na lista). Mas também se considera um contrito positivo o incremento da concentração de respostas certas no topo da lista de resultados, inclusivamente quando a primeira resposta não é melhorada.
As perguntas que são apresentadas nesta secção podem ter origem no QA@CLEF, ou podem ter sido inseridas pelos utilizadores através da interface *Web* do sistema.

4.2.2 Resultados por categoria

O procedimento de avaliação descrito foi aplicado a várias colecções de perguntas, agrupadas por categoria. De seguida apresentam-se listagens com as perguntas em cada conjunto e os valores da análise estatística aos resultados obtidos. Começamos com as categorias de respostas textuais, formadas por factóides ou por expressões mais compridas, passando depois para os géneros de perguntas relativas a locais, quantidades e datas.

QueQual

As perguntas seleccionadas para representar a categoria *QueQual*, na avaliação das técnicas de apreciação de respostas, são apresentadas na primeira coluna da tabela 4.1.

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(3/5) 0,60	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	2	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	não	2	1	(1/5) 0,20	(1/10) 0,10
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	40	1	(0/5) 0,00	(0/10) 0,00
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	não	18	1	(0/5) 0,00	(0/10) 0,00
Que rio nasce na Serra da Estrela?	não	2	1	(2/5) 0,40	(5/10) 0,50
Que prémio recebeu José Saramago?	sim	1	3	(2/5) 0,40	(5/10) 0,50
Que tipo de ave é o milhafre-preto?	não	2	1	(1/5) 0,20	(5/10) 0,50
média	60,00%	5,00	1,80	26,67%	25,11%

Tabela 4.1: QueQual: extracção

Como é característico nesta categoria, são perguntas que pedem algo de determinado tipo explícito: uma constelação, um pássaro e uma espada, para as primeiras questões. Este conjunto de perguntas deu lugar a 316 respostas candidatas, das quais 174 (ou 55%) são consideradas factóides. Todas as respostas foram avaliadas manualmente.

A segunda coluna da tabela indica se a resposta do sistema é correcta ou não. Na terceira e na quarta coluna apresenta-se, respectivamente, o índice da primeira resposta correcta (IPRC) e o índice da primeira resposta errada (IPRE) na lista de resultados de cada pergunta. Depois indica-se o número de resultados considerados correctos entre as cinco primeiras (TOP5), na quinta coluna, e entre as dez primeiras (TOP10) respostas candidatas para cada questão, na sexta coluna.

Para a primeira pergunta, constata-se que a resposta fornecida pelo sistema, que é a primeira da lista de resultados, é uma resposta errada. Pelo valor do IPRC, ficamos a saber que a segunda resposta candidata era correcta.

Pelos valores de TOP5 e TOP10 da mesma pergunta, percebemos que existiam 3 respostas correctas entre as 10 primeiras, que estariam entre as posições 2 e 5. E portanto há uma resposta errada numa das posições: 3, 4 ou 5.
Para algumas perguntas, como a segunda, a terceira e a relativa à nacionalidade de Nicole Kidman, o número de respostas candidatas encontradas pelo sistema pode ser inferior a 10. Nestes casos, o valor usado para a média é a razão entre os acertos e o número de resultados disponíveis, como indicado na respectiva coluna, ao lado da fracção.
A última linha da tabela mostra os valores médios para cada coluna. Em suma, a fase de extracção de respostas nesta categoria, para as questões consideradas, produziu resultados certos em 60% das perguntas. Em média, o IPRC deste conjunto é 5 e o IPRE é 1,80. Idealmente, desejaríamos que IPRC tendesse para 1 e que IPRE fosse maior[3] ou igual a 2. O valor médio do IPRC é prejudicado pelas questões sobre Pearl Harbor e sobre o Cocas, para as quais o primeiro resultado correcto aparece muito abaixo na lista de respostas candidatas. A média de acertos no TOP5 é 26,67%. Na primeira e na terceira pergunta, a concentração de respostas certas entre as cinco primeiras foi a melhor neste teste, com 3 acertos em 5, ou 60%. No outro extremo, temos por exemplo a questão sobre Pearl Harbor, com 0 acertos em 5. O valor médio de TOP10 é 25,11%. Este índice fornece uma informação complementar ao acerto TOP5. Estes valores médios, muito próximos, sugerem uma concentração relativa de respostas correctas aproximadamente igual nos 5 e nos 10 primeiros resultados.

Vistos os resultados obtidos após a fase de extracção, vamos agora observar o efeito da aplicação de uma técnica, na fase seguinte. A tabela 4.2 apresenta os resultados para o mesmo conjunto de perguntas desta categoria (e o mesmo conjunto de respostas), quando se aplica a técnica de bonificação de respostas cujo espectro de activação semântica possui semelhanças com o espectro de outras respostas candidatas, na variante com limite 75 (explicada adiante).
Desta forma, o sistema passa a acertar em 80% dos casos, com mais três perguntas respondidas acertadamente. O acerto médio no TOP5 e no TOP10 sobe. O valor médio para o IPRC desce, o que é positivo, e sobre para o IPRE, igualmente positivo.
Na primeira pergunta não se regista nenhuma melhoria. Antes pelo contrário, pois uma das respostas correctas, que antes estava entre as cinco primeiras, passa a estar mais abaixo na lista ordenada de resultados, situando-se entre os índices 6 e 10. A resposta à segunda pergunta continua a ser correcta, mas nota-se uma pequena melhoria. Desta vez, a segunda resposta na lista também é considerada correcta, com o IPRE a passar para 3.
Na pergunta sobre Pearl Harbor o IPRC passou de 40 para 9, o que significa que já existe uma resposta correcta entre as dez primeiras. A pergunta sobre o Cocas regista também uma melhoria, tendo agora uma resposta correcta no índice 3. As duas últimas perguntas tinham 50% de acertos entre os dez primeiros resultados. O mesmo indicador revela agora 90% e 60%, respectivamente. Em ambas aumentou também o acerto no TOP5, sendo agora 80%.

[3] $IPRE > 2$ acontece sempre que existem várias respostas correctas no topo da lista de resultados

Ao comparar a tabela 4.1 com a tabela 4.2, conclui-se que o sistema obteve melhores resultados com a utilização desta técnica.

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	9	1	(0/5) 0,00	(1/10) 0,10
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	não	3	1	(1/5) 0,20	(1/10) 0,10
Que rio nasce na Serra da Estrela?	sim	1	2	(1/5) 0,20	(6/10) 0,60
Que prémio recebeu José Saramago?	sim	1	3	(4/5) 0,80	(9/10) 0,90
Que tipo de ave é o milhafre-preto?	sim	1	2	(4/5) 0,80	(6/10) 0,60
média	80,00%	1,73	2,07	32,00%	31,11%

Tabela 4.2: QueQual: semelhança de espectro com bonificação limitada a 75

A tabela 4.3 mostra a avaliação dos resultados de outra experiência. Desta vez, após a fase de extracção aplicaram-se duas técnicas: a confirmação de resposta por pesquisa na *Web* e a técnica de semelhança de espectro já mencionada. Como indicam os valores médios, na última linha da tabela, esta combinação de técnicas conduz a resultados um pouco melhores que os anteriores. O sistema respondeu correctamente em 86,67% das perguntas. Há mais uma pergunta respondida acertadamente, relativa ao animal Cocas, que na fase de extracção tinha uma lista de respostas candidatas com a primeira correcta no índice 18. O valor médio do IPRC baixou ligeiramente, sendo prejudicado principalmente pela questão de Pearl Harbor, cujos indicadores não revelam melhorias face à tabela anterior. Confirmando a melhoria, o valor médio do IPRE sobe, denotando a concentração de respostas correctas no topo das lista de resultados de cada pergunta. O

mesmo sugere o aumento do acerto no TOP5. A diminuição no TOP10 é relativamente marginal, pouco relevante quando os outros indicadores melhoram.

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	9	1	(0/5) 0,00	(1/10) 0,10
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio nasce na Serra da Estrela?	sim	1	5	(4/5) 0,80	(5/10) 0,50
Que prémio recebeu José Saramago?	sim	1	5	(4/5) 0,80	(8/10) 0,80
Que tipo de ave é o milhafre-preto?	sim	1	2	(4/5) 0,80	(6/10) 0,60
média	86,67%	1,60	2,47	37,33%	29,78%

Tabela 4.3: QueQual: confirmação *Web* + semelhança de espectro/75

Para concluir a análise de respostas para esta categoria, a tabela 4.4 mostra o resumo com os resultados médios para as diferentes técnicas aplicáveis experimentadas. As técnicas base são:

- reforço por semelhança de espectro de activação de respostas
- reforço por proximidade aos nós mais activos do espectro de activação combinado
- confirmação da resposta por pesquisa na *Web*

Cada técnica segue o seu princípio genérico, referido no capítulo anterior, mas pode, em geral, ser implementada com várias configurações. A tabela 4.4 apresenta resultados para seis variantes da técnica de semelhança de espectro. A primeira delas corresponde

à descrição geral da técnica (secção 3.2.3). Os resultados desta variante são bons para esta categoria, mas detectou-se que o valor de algumas bonificações era bastante elevado. Isto levou às outras variantes, onde existe um limite máximo para o valor que a técnica atribui como bonificação à pontuação base de uma resposta. A ideia subjacente é evitar que termos muito populares (por exemplo: com muitos sinónimos os muitas especializações) beneficiem indevidamente com a abundância de relações na rede semântica. Ao impor um limite máximo à bonificação de cada resposta, a ordem inicial das respostas não é alterada tão radicalmente. A tabela mostra os resultados com os valores médios dos indicadores, para as variantes onde o limite é 15, 30, 45, 75 e 100. Os melhores resultados da primeira técnica surgem quando o limite de bonificação é 75, o que corresponde à variante cujos resultados detalhados foram já relatados na tabela 4.2. Os resultados detalhados por pergunta relativos às restantes variantes desta técnica podem ser consultados nas tabelas A.1 a A.5, no anexo A.

A segunda técnica usada é baseada no espectro semântico combinado, resultante da propagação do nível de activação para todas as respostas. A primeira variante da técnica atribui uma bonificação às respostas cujo nó se encontra perto dos nós com maior nível de activação final na malha de resultados com o espectro combinado. Como tabela indica, com esta técnica o sistema deixa de acertar em muitas das respostas, relativamente à fase de extracção. A taxa de acerto cai de 60% para 26,67%. Observando a tabela A.6 nota-se que para algumas perguntas a primeira resposta correcta foi despromovida, recuando na lista ordenada de resultados. No entanto, os resultados desta variante no que respeita ao número de acertos no TOP5 e no TOP10 não pioram face ao registado na fase de extracção.
As outras variantes da técnica de espectro combinado limitam a bonificação atribuída a cada resposta, tal como para a técnica anterior. Entre estas variantes, o melhor resultado ocorre para o limite de bonificação igual a 15. Apesar de não aumentar a taxa de acerto para a melhor resposta do sistema, comparativamente com a fase de extracção, é nesta variante que se registam os melhores valores nos indicadores IPRC, TOP5 e TOP10.

A terceira técnica usada é a confirmação do valor de resposta através de uma pesquisa na *Web*. A implementação da técnica varia de categoria para categoria. O modo como é formada a expressão a pesquisar (ou as expressões, podem ser várias) é dependente do tipo de pergunta, como previsto na secção 3.2.1. Entre as técnicas individuais, é nesta que se obtém a maior taxa de acerto entre as cinco primeiras respostas, com 37,33%. Comparando com a melhor variante da semelhança de espectro, temos a mesma percentagem de acertos: 80%. Mas as perguntas bem respondidas pelo sistema não são exactamente as mesmas. Comparando as tabelas 4.2 e A.10 podemos observar a diferença, que incide nas respostas sobre o Cocas e sobre o milhafre-preto.

Depois de aplicar individualmente cada técnica, seleccionou-se a melhor variante de cada género, entre aquelas que melhoraram os resultados da fase de extracção. Não se usou a técnica baseada no espectro combinado, por não apresentar ganhos significativos.

		VALORES MÉDIOS			
DESCRIÇÃO DA TÉCNICA	CORRECTA	IPRC	IPRE	Top5	Top10
extracção	60,00%	5,00	1,80	26,67%	25,11%
semelhança de esp. bonif. crescente	73,33%	1,87	2,00	30,67%	30,44%
semelhança de esp. limite 15	80,00%	2,00	2,07	30,67%	29,11%
semelhança de esp. limite 30	80,00%	1,67	2,07	29,33%	29,11%
semelhança de esp. limite 45	80,00%	1,67	2,07	32,00%	29,11%
semelhança de esp. limite 75	80,00%	1,73	2,07	32,00%	31,11%
semelhança de esp. limite 100	73,33%	1,60	1,93	32,00%	30,44%
prox. esp. combinado	26,67%	5,33	1,40	26,67%	25,78%
prox. esp. combinado, limite 15	60,00%	3,67	1,87	29,33%	27,78%
prox. esp. combinado, limite 30	60,00%	3,93	1,93	28,00%	26,44%
prox. esp. combinado, limite 45	53,33%	4,47	1,80	26,67%	25,78%
prox. esp. combinado, limite 75	53,33%	4,80	1,60	25,33%	25,78%
confirmação por pesquisa na Web	80,00%	3,73	2,33	37,33%	29,11%
confirmação Web + sem. de espectro/75	86,67%	1,60	2,47	37,33%	29,78%

Tabela 4.4: QueQual: resumo dos resultados

O sistema aplicou uma combinação de técnicas: confirmação de resposta por pesquisa na *Web* e semelhança de espectro de activação com limite 75. Foi com esta solução que se atingiram os melhores resultados para a categoria *QueQual*, já apresentados em detalhe na tabela 4.3.

Definição

Na categoria *Definição* as respostas são expressões longas, na maior parte dos casos, às quais se podem aplicar várias técnicas de análise. A lista de perguntas escolhidas nesta categoria é apresentada na tabela 4.5. Há questões relativas a conceitos de nome simples, por exemplo *menir*, mas também sobre entidades cuja designação é composta por várias palavras, como *fogo de São Telmo*, ou ainda siglas como *ICCROM*, *RSFSR* ou *VRML*. Para este universo de perguntas, o sistema recolheu 742 respostas candidatas, das quais apenas 112 (15%) são expressões curtas consideradas factóides.

A tabela inclui a avaliação dos resultados da fase de extracção. O sistema respondeu correctamente a 72,73% das perguntas. O valor médio do IPRC é 1,32, o que representa um bom resultado para um sistema de pergunta-resposta. Para reforçar essa leitura, basta observar que todas as perguntas têm pelo menos uma resposta correcta entre no TOP5. A primeira pergunta é uma das que o sistema não respondeu acertadamente. No entanto, a lista de respostas alternativas desta pergunta tem 50% de resultados correctos no TOP10, com três acertos entre as cinco primeiras respostas candidatas.

A fase de extracção encontrou apenas seis respostas candidatas para a questão acerca do *jagertee*. Destas, apenas uma foi classificada como correcta, encontrando-se em 3º

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	não	2	1	(3/5) 0,60	(5/10) 0,50
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	3	(2/5) 0,40	(4/10) 0,40
O que era o A6M Zero?	sim	1	4	(3/5) 0,60	(5/10) 0,50
O que é a paella?	sim	1	4	(3/5) 0,60	(5/10) 0,50
O que é a açorda?	não	2	1	(1/5) 0,20	(5/10) 0,50
O que é o feta?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que é o jagertee?	não	3	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que era a RSFSR?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que é um berimbau?	sim	1	3	(4/5) 0,80	(5/10) 0,50
O que é VRML?	não	2	1	(3/5) 0,60	(6/10) 0,60
O que são os forcados?	não	2	1	(1/5) 0,20	(1/10) 0,10
O que é o fogo de São Telmo?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	2	(3/5) 0,60	(6/10) 0,60
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	4	(3/5) 0,60	(6/10) 0,60
O que é o Crescente Fértil?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(1/5) 0,20	(2/10) 0,20
O que é a Torre do Tombo?	sim	1	2	(1/5) 0,20	(2/10) 0,20
O que são os iaques?	sim	1	3	(3/5) 0,60	(6/10) 0,60
média	72,73%	1,32	2,27	47,27%	41,77%

Tabela 4.5: Definição: extracção

lugar na lista ordenada de resultados para a pergunta.
Os valores médios de acerto entre as cinco e entre as dez primeiras respostas para cada questão são, respectivamente, 47,27% e 41,77%. Estes valores são bastante superiores aos observados na fase de extracção da categoria QueQual. Isto deve-se à natureza da pergunta e ao padrão textual que uma definição pode assumir. A extracção de resultados para as perguntas desta categoria não é tão imprecisa.
Ao propôr uma lista perguntas cujas respostas do sistema têm à partida um elevado grau de acerto, espera-se uma comprovação mais fiável da eficácia das técnicas aplicadas, uma vez que a melhoria acidental dos resultados se torna menos provável.

Aplicando a técnica de bonificação das respostas cujo nó na malha de resultados se encontra próximo aos nós de maior nível de activação do espectro combinado, com o limite máximo igual a 15, os resultados obtidos são os indicados pela tabela 4.6. A técnica melhora o acerto médio da resposta fornecida pelo sistema a cada pergunta, atingindo agora os 90,91%. As duas únicas respostas que o sistema errou foram para as perguntas sobre o *Gil Vicente FC* e a *Torre do Tombo*. Note-se que essas perguntas tinham resposta certa na fase de extracção. Isto evidencia que a reordenação da lista de resultados para cada pergunta, ditada pela técnica aplicada, nem sempre produz uma

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(4/5) 0,80	(4/10) 0,40
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(3/5) 0,60	(6/10) 0,60
O que era o A6M Zero?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que é a paella?	sim	1	2	(3/5) 0,60	(6/10) 0,60
O que é a açorda?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é o feta?	sim	1	8	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que era a RSFSR?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é um berimbau?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é VRML?	sim	1	7	(5/5) 1,00	(6/10) 0,60
O que são os forcados?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	5	(4/5) 0,80	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é o Gil Vicente FC?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	8	1	(0/5) 0,00	(2/10) 0,20
O que são os iaques?	sim	1	4	(4/5) 0,80	(7/10) 0,70
média	90,91%	1,36	3,23	60,91%	49,04%

Tabela 4.6: Definição: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 15

sequência mais correcta que a anterior. Para esta variante da técnica baseada no espectro combinado, pode dizer-se que apesar de errar nestas duas perguntas, o sistema passou a acertar a resposta de seis perguntas onde antes errava. Há portanto um saldo positivo de quatro perguntas correctamente respondidas pelo sistema, o que motiva a subida de 72,73% para 90,91%.
O valor médio do IPRC piorou ligeiramente. Este indicador foi influenciado pela pergunta sobre a *Torre do Tombo*, em que a melhor resposta vem agora no índice 8. Para além da resposta do sistema melhorar, de um modo geral, a ordenação da lista de respostas candidatas do conjunto também melhorou. Pelos valores médios de TOP5 e TOP10 nota-se um aumento de mais de 13% nos acertos entre os cinco primeiros resultados.

A tabela 4.7 apresenta a avaliação dos resultados do sistema com a técnica baseada na frequência de termos. Como foi explicado na secção 3.3.5, esta técnica premeia a afinidade entre resultados pela análise da frequência dos termos que compõem uma resposta no conjunto de todas as respostas candidatas. Na primeira variante, a bonificação é proporcional à frequência média para os termos da resposta. A segunda variante não considera as *stop words*. A tabela mostra os resultados para a terceira variante desta

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(3/5) 0,60	(5/10) 0,50
O que é a Brabançonne?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que era o A6M Zero?	sim	1	7	(5/5) 1,00	(8/10) 0,80
O que é a paella?	sim	1	6	(5/5) 1,00	(5/10) 0,50
O que é a açorda?	sim	1	3	(2/5) 0,40	(6/10) 0,60
O que é o feta?	sim	1	4	(4/5) 0,80	(7/10) 0,70
O que é o jagertee?	não	2	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	7	(5/5) 1,00	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(3/10) 0,30
O que era a RSFSR?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que é um berimbau?	sim	1	3	(4/5) 0,80	(4/10) 0,40
O que é VRML?	sim	1	7	(5/5) 1,00	(8/10) 0,80
O que são os forcados?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é o Crescente Fértil?	sim	1	4	(3/5) 0,60	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(1/5) 0,20	(2/10) 0,20
O que é a Torre do Tombo?	sim	1	2	(1/5) 0,20	(2/10) 0,20
O que são os iaques?	sim	1	6	(5/5) 1,00	(7/10) 0,70
média	95,45%	1,05	3,68	62,73%	47,68%

Tabela 4.7: Definição: frequência de termos 3: filtragem e normalização

técnica, onde a análise de frequência se faz sobre a forma normalizada de cada termo, após a filtragem de *stop words*.
Com esta técnica, apenas uma pergunta foi respondida incorrectamente pelo sistema. Trata-se da pergunta sobre o *jagertee*, que já tinha resposta errada na fase de extracção, mas para a qual a primeira (e única) resposta candidata correcta melhorou, subindo uma posição para o segundo lugar. Nota-se, portanto, uma melhoria no acerto da resposta fornecida em cada pergunta (a primeira da lista), com um excelente valor médio de IPRC: 1,05.
A média de acertos no TOP5 é ligeiramente melhor que a obtida com a técnica anterior. Juntando este facto à leitura dos valores médios IPRC e IPRE nas respectivas tabelas, parece existir uma melhor distribuição de resultados correctos dentro dos cinco primeiros no caso desta técnica baseada na frequência de termos. Isto porque o IPRE médio aumenta.

Como podemos ver na tabela 4.8, as técnicas aplicáveis às respostas nesta categoria de perguntas são:

- reforço por semelhança de espectro de activação de respostas
- reforço por proximidade aos nós mais activos do espectro de activação combinado

	VALORES MÉDIOS				
DESCRIÇÃO DA TÉCNICA	CORRECTA	IPRC	IPRE	Top5	Top10
extracção	72,73%	1,32	2,27	47,27%	41,77%
semelhança de esp. bonif. crescente	72,73%	1,64	3,05	60,00%	50,86%
semelhança de esp. limite 15	86,36%	1,32	3,36	60,91%	49,95%
semelhança de esp. limite 30	81,82%	1,55	3,00	60,91%	51,31%
prox. esp. combinado	81,82%	2,23	2,86	55,45%	46,31%
prox. esp. combinado, limite 15	90,91%	1,36	3,23	60,91%	49,04%
prox. esp. combinado, limite 30	90,91%	1,55	3,41	59,09%	48,59%
prox. esp. combinado, limite 45	86,36%	2,00	2,95	57,27%	46,31%
frequência de termos 1	72,73%	3,77	3,36	59,09%	44,95%
frequência de termos 2	68,18%	3,27	3,86	63,64%	48,13%
frequência de termos 3	95,45%	1,05	3,68	62,73%	47,68%
confirmação por pesquisa na Web 1	77,27%	1,32	2,86	55,45%	46,77%
confirmação por pesquisa na Web 2	81,82%	1,23	3,41	60,91%	48,13%
presença de nomes, min=2, bonif. incremental	90,91%	1,23	3,00	57,27%	47,22%
presença de nomes, min=4, bonif. incremental	77,27%	1,23	3,32	62,73%	49,04%
presença de nomes, taxa	72,73%	1,45	2,27	48,18%	42,22%
presença de nomes, min. no prefixo	86,36%	1,18	2,91	55,45%	45,86%
semelhança de esp. + esp. combinado	86,36%	1,55	3,27	62,73%	49,95%
presença de nomes + esp. combinado	90,91%	1,27	3,36	61,82%	48,13%
pesquisa na Web + esp. combinado	90,91%	1,14	3,77	67,27%	49,04%
pesq. na Web + pr. nomes + esp. comb.	90,91%	1,09	3,77	66,36%	49,95%
freq. de termos + semelhança de esp.	90,91%	1,32	3,77	63,64%	51,31%
freq. de termos + esp. combinado	90,91%	1,18	3,86	67,27%	48,13%
freq. de termos + pesq. na Web	95,45%	1,05	3,95	65,45%	48,59%
freq. de termos + pr. de nomes	90,91%	1,14	3,73	64,55%	51,31%

Tabela 4.8: Definição: resumo dos resultados

- afinidade entre resultados pela frequência de termos
- confirmação da resposta por pesquisa na *Web*
- presença de nomes no texto da resposta

Destas técnicas, aplicadas individualmente, falta apenas mencionar aquela que se baseia na presença de substantivos entre os termos que formam uma resposta. A tabela 4.8 mostra os valores médios para a avaliação das três variantes desta técnica, mencionadas na secção 3.3.5. As duas configurações indicadas para a primeira variante correspondem a exigir pelo menos 2 ou pelo menos 4 nomes, respectivamente, sendo a bonificação gradual com a existência de ainda mais nomes no texto da resposta. Os melhores resultados com esta técnica foram obtidos com a terceira variante, que bonifica as respostas com substantivos no prefixo do texto, e com a primeira versão da primeira variante, que incrementa o acerto médio no TOP5 e no TOP10, ao mesmo tempo que eleva a percentagem de respostas correctas para 90,91%.

Consideremos agora os valores para as três técnicas que também se aplicam à categoria anterior. A confirmação por pesquisa na *Web* continua a melhorar os resultados, em ambas as variantes para esta categoria foi registado uma aumento do número de respostas correctas que o sistema produziu, bem como uma melhoria nos acertos do TOP5 e TOP10.
Tal como já tinha acontecido para a categoria anterior, as variantes com bonificação crescente, sem limite máximo, das técnicas de semelhança de espectro e de espectro combinado, são menos eficazes que as restantes variantes dentro da respectiva técnica. No entanto, nesta categoria todas as variantes baseadas no espectro combinado melhoram os resultados da fase de extracção.
A melhor variante de semelhança de espectro é a que aplica o limite máximo de bonificação igual a 15. Mas inclusivamente a primeira variante tem um efeito positivo nos resultados. Apesar de não melhorar o acerto das respostas fornecidas pelo sistema, consegue melhores valores de acerto nas respostas alternativas, registando-se um aumento nas médias do TOP5 de TOP10.

As últimas oito linhas da tabela 4.8 indicam o resultado da avaliação quando o sistema aplica uma combinação com as melhores variantes de várias técnicas. A primeira dessas experiências combina semelhança de espectro com espectro combinado, obtendo-se um resultado correcto em 86,36% dos casos. Muito embora este valor seja uma melhoria sobre os resultados da fase de extracção, a combinação não permite ao sistema acertar em mais respostas que as conseguidas com qualquer uma das técnicas, isoladamente. Em particular, a variante de espectro combinado com limite 15 conduz o sistema a mais respostas certas. O ganho nesta combinação é um modesto aumento no acerto médio entre o TOP5, ligeiramente acima do obtido em cada uma das suas componentes.
As combinações que maximizam o acerto médio no TOP5, com 67,27% de respostas candidatas correctas nas cinco primeiras do conjunto de resultados de cada pergunta, são as que juntam, por lado pesquisa na *Web* e por outro lado frequência de termos, com o espectro combinado. Mas a melhor resposta que o sistema dá a cada pergunta continua a obter o mesmo grau de acerto que se registava apenas com a técnica de espectro combinado. Nas respostas do sistema, a combinação com o melhor acerto é a junção entre frequência de termos e pesquisa na *Web*, mas também não ultrapassa o valor médio obtido com a terceira variante da técnica baseada na presença de nomes.
Outros resultados, com os valores detalhados por pergunta em cada variante podem ser consultados nas tabelas A.11 a A.32, no anexo A.
Se a soma de duas técnicas não supera o resultado da aplicação de uma só, então estamos na presença de um efeito de cancelamento mútuo entre elas. As respostas que uma técnica beneficia não são necessariamente as escolhidas pela segunda.

Localização

A tabela 4.9 mostra as questões do tipo Localização a cujas respostas se aplicaram as técnicas previstas na secção 3.3.5. São 18 perguntas sobre locais onde se situa alguma

coisa ou em que teve lugar um evento. Enquanto umas questões pedem um local, em sentido lato, outras restringem a natureza desse local, como por exemplo a um distrito, a um país ou a uma ilha. Como a tabela indica, a melhor resposta do sistema para cada pergunta foi considerada correcta para 27,78% dos casos. Nesta colecção, apenas uma pergunta tem menos de dez respostas candidatas na fase de extracção. Com uma reduzida taxa de acerto, só quatro questões possuem mais de 2 resultados correctos entre os 10 primeiros. Em quatro perguntas, a primeira resposta candidata correcta surge fora das 10 primeiras. Dessas, o pior resultado verifica-se na pergunta sobre a *Ossétia do Norte*, onde o IPRC atinge o valor máximo do conjunto: 25. No outro extremo, a per-

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	23	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	não	2	1	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	2	1	(1/5) 0,20	(1/10) 0,10
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Onde é que José Eduardo Pinto da Costa nasceu?	não	3	1	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	13	1	(0/5) 0,00	(0/10) 0,00
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	12	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	4	1	(1/5) 0,20	(5/10) 0,50
Onde fica a Disneyland?	sim	1	2	(3/5) 0,60	(4/10) 0,40
Em que país fica o Muro de Berlim?	não	6	1	(0/5) 0,00	(1/10) 0,10
Onde fica São Bento do Ameixial?	sim	1	7	(5/5) 1,00	(6/10) 0,60
Onde fica Porto Santo?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	27,78%	5,72	1,56	25,56%	19,44%

Tabela 4.9: Localização: extracção

gunta com melhores resultados é a relativa a *São Bento do Ameixial*. A fase de extracção gerou uma sequência de respostas candidatas com 6 acertos concentrados no topo da lista de resultados, respectivamente: *Portugal*, *Estremoz*, *Évora*, *distrito de Évora*, *Europa* e *Alentejo*. Destas resposta, algumas tiveram origem na rede de conhecimento inicial do sistema e em consultas a `maps.google.com`.

Das técnicas baseadas na semântica do valor da resposta, os melhores resultados foram obtidos com a semelhança de espectro de activação, na variante em que a bonificação é limitada a 15. Na tabela 4.10 encontramos os resultados detalhados por pergunta, ao aplicar esta técnica após a extracção.

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	23	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	2	1	(1/5) 0,20	(1/10) 0,10
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Onde é que José Eduardo Pinto da Costa nasceu?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	9	1	(0/5) 0,00	(1/10) 0,10
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	15	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	4	1	(1/5) 0,20	(5/10) 0,50
Onde fica a Disneyland?	sim	1	2	(3/5) 0,60	(4/10) 0,40
Em que país fica o Muro de Berlim?	não	19	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	7	(5/5) 1,00	(6/10) 0,60
Onde fica Porto Santo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	44,44%	6,17	1,83	25,56%	20,00%

Tabela 4.10: Localização: semelhança de espectro com bonificação limitada a 15

Nota-se que não há alteração do número de acertos entre as primeiras cinco e dez respostas, para as primeiras oito perguntas. Ainda assim, a ordenação das respectivas cinco primeiras respostas candidatas não é idêntica, uma vez que o sistema responde correctamente às questões sobre o *Parque Eduardo VII* e sobre o nascimento de *José Eduardo Pinto da Costa*. No total, o sistema acerta em mais 3 questões, subindo para 44,44% de acerto na sua melhor resposta.
Mantém-se o mesmo valor médio de acerto TOP5 e para TOP10 nota-se uma ligeira subida. Apesar do sistema apresentar uma primeira resposta correcta em mais casos, observa-se uma subida do IPRC médio para 6,17. Isto significa que algumas das per-

guntas não respondidas tiveram um agravamento na sua lista de resultados. O valor médio do IPRE subiu, o que é positivo e usual quando se mantém o acerto TOP5 e simultaneamente melhora o acerto na primeira resposta.

Outra técnica aplicável é a confirmação do valor de cada resposta candidata com uma pesquisa na *Web*, que relaciona esse valor com o alvo da questão. Como indica a tabela 4.11, o sistema responde correctamente a 44,44% das perguntas, o que corresponde à mesma percentagem que se tinha registado com a técnica anterior. Importa salientar que as oito perguntas com resposta correcta não são as mesmas em ambas as técnicas.

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	8	1	(0/5) 0,00	(1/10) 0,10
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	não	2	1	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Onde é que José Eduardo Pinto da Costa nasceu?	não	4	1	(1/5) 0,20	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	13	1	(0/5) 0,00	(0/10) 0,00
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	14	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	4	1	(1/5) 0,20	(5/10) 0,50
Onde fica a Disneyland?	sim	1	3	(2/5) 0,40	(4/10) 0,40
Em que país fica o Muro de Berlim?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Onde fica São Bento do Ameixial?	sim	1	7	(5/5) 1,00	(6/10) 0,60
Onde fica Porto Santo?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	44,44%	4,67	1,78	24,44%	20,00%

Tabela 4.11: Localização: confirmação por pesquisa *Web*

Dois exemplos disso são as perguntas sobre o *Parque Eduardo VII* e sobre o *Muro de Berlim*, cada uma acertada apenas por uma das técnicas. Estas diferenças, inerentes à natureza de cada técnica, sugerem uma combinação entre as técnicas. O resultado da avaliação de tal combinação é apresentado mais adiante e a tabela A.41 tem esses resultados detalhados por pergunta.

Relativamente à fase de extracção, a técnica baseada na pesquisa *Web* apresenta valores de acerto médio TOP5 e TOP10 aproximadamente iguais (tal como na técnica anterior). O valor médio do IPRC é mais favorável, descendo para 4,67. Há menos perguntas sem respostas candidatas correctas entre as dez primeiras.

O resumo dos resultados para as técnicas e combinações de técnicas ensaiadas nesta categoria é indicado pela tabela 4.12. Mais uma vez, as variantes sem limite de bonificação por resposta, da técnica baseada na semelhança de espectro de activação semântica e da técnica de espectro combinado, apresentam um acerto médio reduzido, chegando mesmo a piorar os resultados da fase de extracção entre as cinco e entre as dez primeiras respostas.
A combinação das melhores variantes, já apresentadas, mantém 44,44% de acerto na primeira resposta, idêntico ao conseguido com cada uma das técnicas que aplica. Mas para os restantes indicadores, foi com esta combinação que se registaram os melhores valores médios: IPRC mínimo, IPRE máximo, TOP10 máximo e TOP5 igual à fase de extracção e igualmente o maior observado.
As duas combinações restantes, semelhança de espectro com espectro combinado e pesquisa *Web* com espectro combinado, levaram a um acerto melhor que o obtido na extracção, mas que fica aquém do registado por qualquer das técnicas em separado. Os resultados detalhados destas combinações encontram-se nas tabelas A.40 e A.42, respectivamente.

		VALORES MÉDIOS			
DESCRIÇÃO DA TÉCNICA	CORRECTA	IPRC	IPRE	Top5	Top10
extracção	27,78%	5,72	1,56	25,56%	19,44%
semelhança de esp. bonif. crescente	27,78%	13,67	1,44	12,22%	12,22%
semelhança de esp. limite 15	44,44%	6,17	1,83	25,56%	20,00%
semelhança de esp. limite 30	38,89%	8,33	1,72	24,44%	18,89%
semelhança de esp. limite 45	33,33%	9,83	1,67	23,33%	16,67%
prox. esp. combinado	27,78%	8,72	1,44	13,33%	12,22%
prox. esp. combinado, limite 15	38,89%	6,44	1,72	22,22%	16,67%
prox. esp. combinado, limite 30	38,89%	7,56	1,61	20,00%	13,33%
prox. esp. combinado, limite 45	38,89%	8,22	1,56	16,67%	12,78%
confirmação por pesquisa na Web	44,44%	4,67	1,78	24,44%	20,00%
sem. de espectro + esp. combinado	33,33%	8,06	1,50	21,11%	16,67%
pesquisa na Web + sem. de espectro	44,44%	4,61	1,83	25,56%	21,11%
pesquisa na Web + esp. combinado	33,33%	5,06	1,56	25,56%	18,33%

Tabela 4.12: Localização: resumo dos resultados

Quantidade

As respostas desta categoria são quantidades. Independentemente do formato que apresentam, elas têm um significado avaliável. Ao encontrar uma lista de resultados com alguns valores dispersos e outros muito concentrados, próximos a determinado valor, o sistema pode tomar como mais credíveis os resultados aglomerados. Isto corresponde à técnica de bonificação de respostas com significado semelhante, só que para este caso não há uma malha de nós e não se aplica a propagação do nível de activação, pois a semelhança semântica entre respostas detecta-se directamente.
As perguntas escolhidas para esta categoria estão indicadas na tabela 4.13. São 16 questões sobre quantidades de natureza distinta, algumas enquadradas nas subclasses de *Quantidade* apresentadas na secção 3.1.5, especializadas em idades, distâncias ou valores monetários.

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quantos jogadores tem um time de basquete?	não	104	1	(0/5) 0,00	(0/10) 0,00
Qual o diâmetro de Ceres?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Quantos focos tem uma elipse?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que profundidade atinge a Fossa das Marianas?	sim	1	2	(2/5) 0,40	(6/10) 0,60
Qual a população da Ilha de Moçambique?	não	13	1	(0/5) 0,00	(0/10) 0,00
Quanto custou a Mars Observer?	sim	1	3	(2/3) 0,67	(2/3) 0,67
Qual a altura do Kebnekaise?	sim	1	4	(3/5) 0,60	(5/8) 0,62
Quantos ossos têm a face?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Com que idade o Mequinho foi campeão brasileiro de xadrez?	não	2	1	(2/5) 0,40	(2/6) 0,33
Quantos habitantes tinha Berlim em 1850?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Qual é a temperatura do zero absoluto?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Quantas faixas tem a bandeira dos Estados Unidos?	não	2	1	(1/5) 0,20	(1/10) 0,10
Quantos filmes realizou Jean Vigo?	não	2	1	(1/5) 0,20	(1/10) 0,10
Qual o comprimento da Ponte do Øresund?	sim	1	3	(3/5) 0,60	(4/10) 0,40
Quantas esposas tinha Ngungunhane?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Quantos metros saltou Nelson Évora nos Jogos Olímpicos?	sim	1	2	(2/5) 0,40	(2/8) 0,25
média	68,75%	8,38	1,94	34,17%	29,84%

Tabela 4.13: Quantidade: extracção

A fase de extracção de respostas para estas questões levou a um resultado correcto

em 68,75% dos casos, com um acerto entre as primeiras cinco e entre as primeiras dez respostas candidatas de 34,17% e 29,84%, respectivamente. A primeira pergunta da lista, redigida em Português do Brasil e proveniente do QA@CLEF2007, tem uma lista de resultados muito extensa e formada por valores errados até ao índice 104, como indica o IPRC. Só cinco perguntas registam mais de 2 acertos no TOP10. No outro extremo, há duas perguntas sem qualquer acerto nos primeiros dez resultados.

Como descrito na secção 3.3.5, as técnicas aplicáveis a esta categoria de perguntas sobre quantidades são:

- afinidade semântica
- confirmação da resposta por pesquisa na *Web*

A primeira variante da técnica de análise da afinidade semântica entre resultados faz a atribuição de pontos de bonificação a uma resposta com base na existência de outros resultados com valor próximo. Um resultado próximo à resposta estará numa vizinhança do ponto que a representa no eixo semântico do espaço multidimensional. A figura 3.6 ilustra a proximidade existente entre várias respostas, com valores discretos e com intervalos. A análise da proximidade só se aplica se não houver incompatibilidade entre eventuais unidades de medida associadas a cada valor.
A segunda variante segue o mesmo princípio de bonificação, mas considerando apenas os resultados tidos como mais plausíveis. Dito de outro modo, uma resposta é valorizada apenas se existir na sua vizinhança um ou mais resultados com pontuação base acima da média. A pontuação base é a resultante da fase de extracção.

A tabela 4.14 mostra os resultados obtidos quando se aplica a segunda variante da técnica de afinidade semântica entre quantidades, para a mesma lista de perguntas. A percentagem de perguntas respondidas correctamente pelo sistema sobre para 75%, por haver agora mais uma primeira resposta correcta, relativa ao *campeão brasileiro de xadrez*. Apesar deste ganho, o valor médio do IPRC piora, subindo para 8,56, devido à primeira pergunta. Os valores médios dos restantes indicadores revelam uma pequena melhoria. O IPRE melhora, subindo para 2,44. O número de acertos nas primeiras respostas também é incrementado, em particular nos cinco primeiros resultados, com TOP5 igual a 40,42%.
Para além da questão que passou a ter resposta correcta, também beneficiaram com a aplicação desta técnica as perguntas relativas à *Fossa das Marianas* e ao monte *Kebnekaise*. A primeira destas, passou a ter 6 resultados correctos nos primeiros índices da sua lista de resultados, como indica o respectivo IPRE.

A técnica baseada na pesquisa do valor de cada resposta candidata na *Web* tem uma particularidade relevante nesta categoria de perguntas. No nível abstracto, a técnica funciona do mesmo modo, realizando pesquisas na *Web* com hipóteses que resultam da junção de partes da pergunta com a resposta. A novidade tem a ver com o valor

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quantos jogadores tem um time de basquete?	não	108	1	(0/5) 0,00	(0/10) 0,00
Qual o diâmetro de Ceres?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Quantos focos tem uma elipse?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que profundidade atinge a Fossa das Marianas?	sim	1	7	(5/5) 1,00	(7/10) 0,70
Qual a população da Ilha de Moçambique?	não	13	1	(0/5) 0,00	(0/10) 0,00
Quanto custou a Mars Observer?	sim	1	3	(2/3) 0,67	(2/3) 0,67
Qual a altura do Kebnekaise?	sim	1	6	(5/5) 1,00	(5/8) 0,62
Quantos ossos têm a face?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Com que idade o Mequinho foi campeão brasileiro de xadrez?	sim	1	2	(2/5) 0,40	(2/6) 0,33
Quantos habitantes tinha Berlim em 1850?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Qual é a temperatura do zero absoluto?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Quantas faixas tem a bandeira dos Estados Unidos?	não	2	1	(1/5) 0,20	(1/10) 0,10
Quantos filmes realizou Jean Vigo?	não	2	1	(1/5) 0,20	(1/10) 0,10
Qual o comprimento da Ponte do Øresund?	sim	1	3	(3/5) 0,60	(4/10) 0,40
Quantas esposas tinha Ngungunhane?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Quantos metros saltou Nelson Évora nos Jogos Olímpicos?	sim	1	2	(2/5) 0,40	(2/8) 0,25
média	75,00%	8,56	2,44	40,42%	30,47%

Tabela 4.14: Quantidade: afinidade semântica 2

da resposta. Neste tipo de perguntas, os resultados numéricos são convertidos do formato original (por extenso ou com algarismos organizados de diferentes maneiras) para o formato normalizado do sistema, que é baseado em algarismos juntamente com os separadores usuais em Portugal. A técnica efectua a pesquisa de cada hipótese com ambas as modalidades para o valor da resposta, o formato normalizado e o formato textual original.
Na tabela A.44 podemos consultar a avaliação dos resultados obtidos com a aplicação desta técnica sobre o mesmo conjunto de perguntas. Resumidamente, há a destacar apenas uma ligeira subida dos valores médios do IPRE e do acerto nas primeiras respostas candidatas, relativamente à fase de extracção.

A tabela 4.15 apresenta os valores médios resultantes de cada variante testada neste tipo de pergunta. Uma combinação de técnicas, que foi igualmente testada, junta a se-

		VALORES MÉDIOS			
DESCRIÇÃO DA TÉCNICA	CORRECTA	IPRC	IPRE	Top5	Top10
extracção	68,75%	8,38	1,94	34,17%	29,84%
afinidade semântica 1	68,75%	8,44	2,38	39,17%	29,84%
afinidade semântica 2	75,00%	8,56	2,44	40,42%	30,47%
confirmação por pesquisa na Web	68,75%	8,38	2,25	37,92%	31,09%
afinidade + pesquisa na Web	75,00%	8,56	2,62	40,42%	31,72%

Tabela 4.15: Quantidade: resumo dos resultados

gunda variante da técnica baseada na afinidade entre respostas e a pesquisa na *Web*. A combinação melhora os resultados da fase de extracção, mas não demonstrou vantagens relevantes face à aplicação isolada da técnica de afinidade semântica.

Quando

A lista de perguntas seleccionada para avaliar os resultados do sistema na categoria *Quando* é apresentada na tabela 4.16. São 19 questões sobre a data de determinado acontecimento, para as quais foram extraídas 551 respostas candidatas, ao todo.
A tabela mostra a avaliação dos resultados da fase de extracção. O sistema responde correctamente a apenas 36,84% das perguntas do conjunto, o que corresponde a 7 questões. Existe um reduzido número de respostas candidatas correctas. Se analisarmos a coluna da direita, verificamos que apenas 3 das 19 perguntas têm mais de 2 respostas correctas entre as dez primeiras, e nessas, o valor é 3. O valor médio no TOP10 é 16,97%.
Neste conjunto, a diferença relativa entre TOP5 e TOP10 é considerável: o acerto médio nas 10 primeiras respostas equivale a aproximadamente 64% do acerto médio nas 5 primeiras. Isto sugere que as respostas certas no TOP10 se situam maioritariamente entre as cinco primeiras.
O indicador IPRE assume o valor 1 na maior parte das questões, não indo além do valor 3. É, portanto, um conjunto de resultados com muito espaço para melhorias.

As respostas com datas podem ser organizadas de acordo com o respectivo significado, tal como sugere a figura 3.5 relativamente ao eixo semântico do espaço de resultados. As respostas de unidade temporal mais abrangente, como ano ou século, correspondem a intervalos no eixo do tempo. Esta abordagem simplifica a quantificação da distância semântica entre duas respostas. Como consequência, as técnicas usadas para o tipo *Quantidade* podem aplicar-se também nesta categoria:

- afinidade semântica
- confirmação da resposta por pesquisa na *Web*

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quando foi fundado o Nacional da Madeira?	não	3	1	(1/5) 0,20	(1/10) 0,10
Quando foi registado pela primeira vez o cometa Halley?	não	7	1	(0/5) 0,00	(1/10) 0,10
Quando se realizaram os Jogos Olímpicos de Munique?	não	5	1	(1/5) 0,20	(1/10) 0,10
Em que dia ocorreu a batalha de La Lys?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Quando é que foi assinada a Lei Áurea?	sim	1	2	(2/5) 0,40	(2/5) 0,40
Em que ano foi fundada a Bertrand?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Quando nasceu Machado de Assis?	não	3	1	(2/5) 0,40	(2/10) 0,20
Quando foi inaugurado o metropolitano de Lisboa?	não	3	1	(1/5) 0,20	(1/10) 0,10
Quando começa o outono no hemisfério sul?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que período existiu a República de Veneza?	não	10	1	(0/5) 0,00	(1/10) 0,10
Em que ano houve um terramoto no Irão?	não	18	1	(0/5) 0,00	(0/10) 0,00
Quando começou o Neolítico?	não	2	1	(2/5) 0,40	(2/10) 0,20
Quando nasceu Thomas Mann?	sim	1	2	(2/5) 0,40	(3/10) 0,30
Quando reinou Isabel II de Castela?	não	32	1	(0/5) 0,00	(0/10) 0,00
Em que ano foi construída a sinagoga de Curação?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Quando foi assinado o Tratado de Zamora?	sim	1	3	(2/5) 0,40	(3/10) 0,30
Quando é que viveu Zenão de Eleia?	não	2	1	(1/5) 0,20	(1/8) 0,12
Quando foi fundado o Vasco da Gama?	não	3	1	(1/5) 0,20	(2/10) 0,20
Quando nasceu Vasco da Gama?	sim	1	3	(2/5) 0,40	(3/10) 0,30
média	36,84%	5,11	1,53	25,26%	16,97%

Tabela 4.16: Quando: extracção

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quando foi fundado o Nacional da Madeira?	não	2	1	(1/5) 0,20	(1/10) 0,10
Quando foi registado pela primeira vez o cometa Halley?	não	7	1	(0/5) 0,00	(1/10) 0,10
Quando se realizaram os Jogos Olímpicos de Munique?	não	5	1	(1/5) 0,20	(1/10) 0,10
Em que dia ocorreu a batalha de La Lys?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Quando é que foi assinada a Lei Áurea?	sim	1	2	(2/5) 0,40	(2/5) 0,40
Em que ano foi fundada a Bertrand?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Quando nasceu Machado de Assis?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Quando foi inaugurado o metropolitano de Lisboa?	não	5	1	(1/5) 0,20	(1/10) 0,10
Quando começa o outono no hemisfério sul?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que período existiu a República de Veneza?	não	8	1	(0/5) 0,00	(1/10) 0,10
Em que ano houve um terramoto no Irão?	não	18	1	(0/5) 0,00	(0/10) 0,00
Quando começou o Neolítico?	não	2	1	(2/5) 0,40	(2/10) 0,20
Quando nasceu Thomas Mann?	sim	1	2	(2/5) 0,40	(3/10) 0,30
Quando reinou Isabel II de Castela?	não	34	1	(0/5) 0,00	(0/10) 0,00
Em que ano foi construída a sinagoga de Curação?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Quando foi assinado o Tratado de Zamora?	sim	1	3	(2/5) 0,40	(3/10) 0,30
Quando é que viveu Zenão de Eleia?	não	2	1	(1/5) 0,20	(1/8) 0,12
Quando foi fundado o Vasco da Gama?	não	3	1	(1/5) 0,20	(1/10) 0,10
Quando nasceu Vasco da Gama?	sim	1	3	(2/5) 0,40	(3/10) 0,30
média	42,11%	5,05	1,58	25,26%	16,45%

Tabela 4.17: Quando: afinidade semântica 2

A técnica que, aplicada individualmente, conduz aos melhores resultados é a afinidade semântica, na sua segunda variante. Como indica a tabela 4.17, a taxa de acerto do sistema sobe para 42,11%, porque a técnica leva o sistema a acertar também na data do nascimento de *Machado de Assis*. O valor médio do acerto no TOP5 é exactamente igual ao registado na fase de extracção. Para os restantes indicadores, a diferença relativamente à fase de extracção é mínima. O sentido em que evoluem IPRC e IPRE (descendo e subindo, respectivamente) é positivo, mas numa escala demasiado pequena para se registar uma melhoria relevante.

A técnica de pesquisa na *Web* é também aplicável para os resultados desta categoria. Em cada resposta, a pesquisa faz-se com o valor normalizado da data (exemplo: 2009-11-01, para o Dia da Universidade de Évora) e também com o respectivo valor textual original (1 de Novembro de 2009). A tabela A.47 tem os resultados detalhados para esta técnica, que responde bem ao mesmo número de perguntas, compensando um erro na data da *batalha de La Lys* com um acerto nos *Jogos Olímpicos de Munique*.

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quando foi fundado o Nacional da Madeira?	não	4	1	(1/5) 0,20	(1/10) 0,10
Quando foi registado pela primeira vez o cometa Halley?	não	29	1	(0/5) 0,00	(0/10) 0,00
Quando se realizaram os Jogos Olímpicos de Munique?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que dia ocorreu a batalha de La Lys?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Quando é que foi assinada a Lei Áurea?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Em que ano foi fundada a Bertrand?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Quando nasceu Machado de Assis?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Quando foi inaugurado o metropolitano de Lisboa?	não	4	1	(1/5) 0,20	(1/10) 0,10
Quando começa o outono no hemisfério sul?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que período existiu a República de Veneza?	não	9	1	(0/5) 0,00	(1/10) 0,10
Em que ano houve um terramoto no Irão?	não	18	1	(0/5) 0,00	(0/10) 0,00
Quando começou o Neolítico?	não	2	1	(2/5) 0,40	(2/10) 0,20
Quando nasceu Thomas Mann?	sim	1	2	(1/5) 0,20	(3/10) 0,30
Quando reinou Isabel II de Castela?	não	34	1	(0/5) 0,00	(0/10) 0,00
Em que ano foi construída a sinagoga de Curação?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Quando foi assinado o Tratado de Zamora?	sim	1	2	(2/5) 0,40	(3/10) 0,30
Quando é que viveu Zenão de Eleia?	não	2	1	(1/5) 0,20	(1/8) 0,12
Quando foi fundado o Vasco da Gama?	não	4	1	(1/5) 0,20	(1/10) 0,10
Quando nasceu Vasco da Gama?	sim	1	3	(2/5) 0,40	(3/10) 0,30
média	47,37%	6,16	1,63	24,21%	15,92%

Tabela 4.18: Quando: afinidade + pesquisa *Web*

A tabela 4.18 apresenta os resultados da combinação entre a técnica anterior e a confirmação do valor de resposta por pesquisa na *Web*. Relativamente à tabela anterior, da análise de afinidade semântica, salienta-se a subida da percentagem de acerto para os 47,37%, com a pergunta sobre os *Jogos Olímpicos de Munique* a ser agora correctamente respondida pelo sistema. O valor médio do IPRC sobe para 6,16, prejudicado pelos resultados da pergunta sobre o *cometa Halley*, onde a componente de pesquisa na *Web* promoveu bastantes respostas erradas. Os valores médios nesta combinação de técnicas, para o acerto entre as primeiras respostas candidatas são próximos aos registados na fase de extracção, ainda que um pouco inferiores. O IPRE médio é igual ao registado apenas com a técnica de pesquisa na *Web*, que é o mais elevado nesta categoria, o que significa uma melhoria face aos resultados iniciais.

Para terminar a análise da avaliação para esta categoria, a tabela 4.19 tem uma síntese dos resultados por cada variante ou combinação de técnicas usadas.

		VALORES MÉDIOS			
DESCRIÇÃO DA TÉCNICA	CORRECTA	IPRC	IPRE	Top5	Top10
extracção	36,84%	5,11	1,53	25,26%	16,97%
afinidade semântica 1	36,84%	5,74	1,42	25,26%	15,39%
afinidade semântica 2	42,11%	5,05	1,58	25,26%	16,45%
confirmação por pesquisa na Web	42,11%	5,63	1,63	22,11%	15,92%
afinidade + pesquisa na Web	47,37%	6,16	1,63	24,21%	15,92%

Tabela 4.19: Quando: resumo dos resultados

4.3 Resumo

As secções anteriores descrevem algumas experiências realizadas com o sistema SENSO, desenvolvido segundo o modelo teórico apresentado. Essas experiências envolveram:

- simulações para testar separadamente alguns módulos constituintes do sistema
- sessões de demonstração em que o sistema foi sujeito à utilização por pessoas, processo que ajudou na identificação de aspectos a corrigir
- participação em duas edições da actividade QA@CLEF
- avaliação da fase de extracção do sistema
- avaliação do contributo de cada técnica de apreciação de respostas candidatas para os resultados do sistema

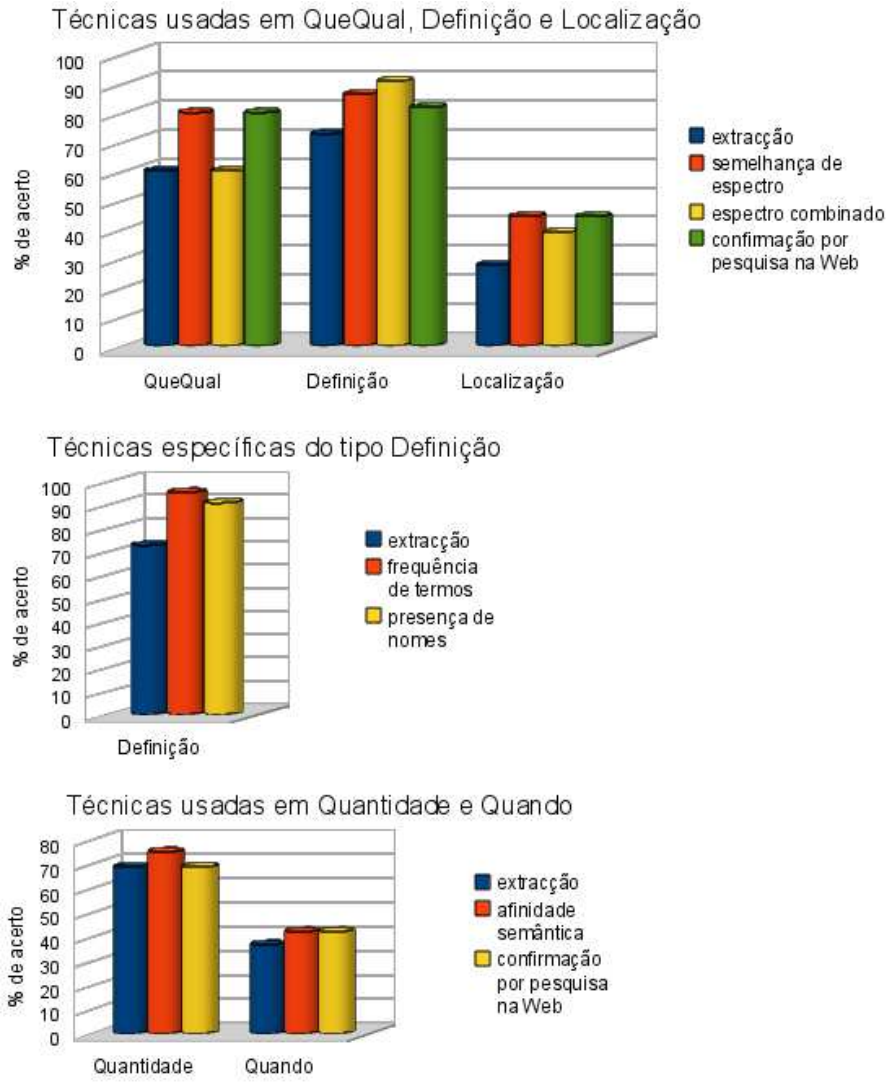

Figura 4.2: Síntese com a percentagem de perguntas acertadas com a melhor variante de cada técnica, em comparação com a fase de extracção

Na figura 4.2 podemos observar a evolução da percentagem de acerto do sistema quando se utiliza determinada técnica de reordenação de respostas, após a fase de extracção. A figura tem três gráficos. Cada gráfico tem uma coluna azul, do lado esquerdo de cada categoria de perguntas, seguida de uma coluna para cada técnica aplicada. Não estão representadas as combinações de técnicas. Para cada técnica foi usado o resultado decorrente da sua melhor variante. Da observação dos gráficos resulta que a aplicação das técnicas melhora os resultados da fase de extracção. Essa melhoria é mais expressiva nas categorias *QueQual*, *Definição* e *Localização*. O terceiro gráfico evidencia um progresso menos significativo nas categorias *Quantidade* e *Quando*, tanto para a técnica baseada

na afinidade semântica como para a técnica baseada na confirmação por pesquisa na *Web*.
O próximo capítulo tem as conclusões desta dissertação, onde se inclui a apreciação do trabalho com uma análise crítica dos resultados obtidos e que aqui foram apresentados.

Capítulo 5

Conclusões

Este capítulo conclui o documento com a memória desta Tese. A primeira secção recapitula os aspectos principais no modelo teórico proposto. Em 5.2 estabelece-se uma comparação entre o sistema SENSO e outros SPR referidos no capítulo 2. A secção 5.3 é dedicada à apreciação dos resultados da avaliação do sistema, fazendo também um balanço com as contribuições deste trabalho. Para terminar a dissertação, a secção 5.4 enumera possibilidades para trabalho futuro, dando continuidade ao que aqui se expõe.

5.1 O Modelo

Nesta dissertação é proposto um modelo teórico para um Sistema de Pergunta-Resposta de âmbito geral que, em termos abstractos, usa como referência o comportamento humano ao decidir por uma resposta entre um conjunto de possíveis resultados. O aspecto central do modelo é a contextualização de respostas, não apenas no tempo e no espaço, mas também na vertente semântica. O espaço multidimensional de respostas é a estrutura do modelo onde os resultados provisórios são enquadrados nos eixos temporal, espacial e semântico, como foi descrito na secção 3.1.6. Esta organização permite a diferenciação das respostas candidatas, sob várias perspectivas, o que pode ser explorado pelas técnicas de ordenação e escolha de respostas.
As perguntas, que podem incidir sobre qualquer assunto, são inicialmente classificadas numa categoria, como *Definição*, *Localização* ou *Quantidade*. Cada categoria tem um determinado género de resposta e formas específicas de localização dessas respostas. A eventual escolha de algumas fontes de informação especializadas é também determinada pela categoria de questão. Por conseguinte, a classificação é uma parte importante do processo e fornece uma orientação para a procura da resposta.

O modelo inclui uma rede semântica como base de conhecimento inicial. Esta base de conhecimento ajuda na interpretação de significados possíveis para termos em respostas ou frases encontradas durante o processamento de uma pergunta. Pode conter directamente a resposta para algumas questões, ou ainda auxiliar na identificação de outras

respostas, pela análise e associação a outros elementos de informação. É uma fonte de informação estruturada que, tal como qualquer ontologia que o sistema possa consultar, fornece indicações semânticas sobre conceitos. Na contextualização semântica, estas indicações são pistas para escolher ou para descartar uma possível resposta.
Outras fontes de informação consultadas são os dicionários e enciclopédias *online*, cujo conteúdo é considerado fidedigno e particularmente útil para as definições. Como exemplos de fontes especializadas, temos ainda os recursos geográficos, como a GEO-NET-PT ou o serviço `maps.google.com`. Os documentos com texto livre são uma fonte de informação não estruturada e existem em abundância no formato digital. Sejam colecções locais de textos ou documentos espalhados por todo o mundo e acessíveis via *Web*, estes repositórios são fonte de respostas para o modelo.
Pelo volume de dados que contém, pela sua abrangência temática, dinamismo e facilidade de acesso, a *Web* é um recurso incontornável que pessoas de todas as idades se habituaram a usar para as suas necessidades de informação. O habitual procedimento humano de procura de documentos e análise do respectivo conteúdo pode ser automatizado e integrado na metodologia de um SPR. Um dos aspectos chave neste procedimento é a definição da consulta para a procura de documentos. O outro aspecto essencial é a capacidade de encontrar uma resposta no conteúdo desses documentos. É sobretudo neste último aspecto que os SPR ficam aquém do ser humano. O processo automático de identificação de respostas em textos da *Web*, neste modelo, assenta na procura superficial de padrões no texto, podendo ser complementada com uma análise linguística localizada. Tirando partido da redundância existente na *Web*, este processo permite a captação de resultados que precisam depois de uma apreciação suplementar.

Cada resposta candidata tem uma pontuação que reflecte o respectivo grau de confiança. Este valor resulta da precisão estimada para a regra de extracção usada, das características da frase de origem da resposta e da fiabilidade da fonte de informação. A base de conhecimento do sistema e os dicionários têm maior credibilidade que textos exteriores ao sistema. Em determinados domínios, as respostas de uma fonte de informação estruturada e especializada no tema da pergunta têm grau de confiança superior ao atribuído a respostas obtidas em textos diversos.
Esta graduação inicial das respostas candidatas resulta de uma apreciação individual de cada resultado durante a fase de extracção. O modelo inclui uma fase posterior para avaliação de cada resposta candidata no âmbito da pergunta e considerando também as demais alternativas. A nível individual, a confirmação de uma resposta pela consulta de outras fontes de informação é usada para incrementar a pontuação inicial. Depois, a contextualização de cada resultado, no tempo, no espaço e em termos semânticos pode ser avaliada face ao previsto para a pergunta. Se a categoria da questão prevê respostas de um tipo conceptual específico, então os resultados explicitamente inadequados no plano semântico podem ser descartados. Do mesmo modo, se os contextos espacial e temporal de um resultado são determinados e incompatíveis com o pretendido na questão, então esse resultado deve ser preterido, para que restem apenas resultados válidos para a época e espaço a que a pergunta se reporta.

Para a apreciação sobre o colectivo, consideram-se todos os resultados captados como possíveis respostas. O modelo valoriza respostas que se julgam mais coerentes, que sejam reforçadas por outros resultados com os quais possuam afinidade de significado. A avaliação desta afinidade semântica entre conceitos tem por base modelos sobre a recuperação de informação na memória humana. Nestes modelos com inspiração nas neurociências, a memória é vista como uma estrutura associativa onde ocorre activação de elementos de informação. Através das relações semânticas, o nível de activação de um elemento é propagado para outros elementos relacionados.
O modelo propõe duas técnicas para detectar eventuais tendências semânticas no conjunto dos resultados possíveis e valorizar respostas que encontrem reforço nesse enquadramento semântico. A metodologia da primeira técnica consiste em determinar o espectro de activação semântica de cada resposta candidata e bonificar as que tiverem semelhanças no respectivo espectro. A segunda técnica proposta faz a activação semântica de todas as respostas e valoriza resultados próximos aos nós de maior activação no espectro combinado.
Em termos abstractos, esta metodologia é independente do tema da pergunta e também do idioma usado na pergunta ou nos textos consultados nas diversas fontes de informação. As técnicas propostas também não são específicas de um género de pergunta em particular. No âmbito deste trabalho, as técnicas foram aplicadas ao espaço de resultados de questões com a categoria *QueQual*, *Definição*, *Localização*, *Quantidade* e *Quando*, tendo sido avaliado o respectivo efeito.

5.2 Comparação do sistema com outros SPR

O sistema SENSO é um SPR implementado com a estrutura modular prevista no modelo teórico proposto e permitiu a aplicação das técnicas propostas como objectivo específico desta investigação. A tabela 5.1 faz uma síntese das principais características do sistema SENSO, juntamente com outros SPR já mencionados na secção 2.3. Todos estes sistemas aceitam perguntas em Português e procuram resposta em fontes na mesma língua.
Na primeira coluna observamos que a classificação da pergunta é prática comum, servindo para orientar a procura de respostas em função do género e foco da questão. As categorias em que a pergunta pode ser classificada variam de sistema para sistema, em quantidade e grau de especialização. O sistema PRIBERAM usa as categorias de pergunta também na fase de pré-processamento dos textos, assinalando as frases relevantes para determinado género de pergunta.
Dos oito sistemas descritos, quatro têm como fonte de respostas colecções locais de documentos, enquanto que a outra metade recorre também à *Web*. No sistema SENSO esses dois tipos de fonte aplicam-se a qualquer questão, e outras fontes, que podem ser estruturadas ou mais especializadas, podem também ser consultadas na procura de respostas, em função da categoria de pergunta.
O reconhecimento de entidades mencionadas, a selecção de trechos e a análise morfo-sintáctica são procedimentos explicitamente seguidos pela maioria.

sistema	pergunta	fontes	extracção	técnicas/ferramentas
PERGUNTE	Classificação da pergunta em categorias.	*Web*. Recuperação de documentos pelos motores de busca da *Web*.	procura superficial de padrões no texto	agrupa respostas semanticamente semelhantes
XISQUÊ	análise da questão: verbo principal e termos relevantes	*Web*. Recuperação de documentos pelos motores de busca da *Web*. ST selecciona frases daqueles documentos.	*resposta curta*: nome ou expressão da frase escolhida com base em regras. *resposta longa*: frase resultante de ST, nos caso em que não há *resposta curta*.	REM. Requisito: responder em tempo real. testes indicam tempo médio de resposta: 21 seg.
RAPOSA	Classificação da pergunta em categorias.	Colecções de textos. Expansão da expressão de pesquisa para recuperação de documentos e ST.	Dois procedimentos: procura superficial de padrões no texto e análise estatística de expressões compatíveis com o tipo de pergunta.	Fusão de respostas candidatas por agrupamento de expressões de igual significado. Analisador JSPELL e REM com SIEMÊS.
ESFINGE	Resolução de anáfora. Geram-se padrões de pesquisa a partir da questão.	*Web* e textos locais. ST com várias iterações cada vez menos restritivas.	Procura superficial de padrões no texto + entidades mencionadas + selecção de n-gramas. Filtragem de resultados. Ordenação recorre ao repositório de co-ocorrências BACO.	Resolução de anáfora na pergunta, em função das anteriores. REM com SIEMÊS. Analisador PALAVRAS. Interface *Web*.
QA@L2F	Classificação da pergunta, e detecção do foco e elementos chave.	Colecções de textos. Processamento linguístico sobre os textos.	Por detecção de entidades mencionadas próximas ao foco da questão e por procura superficial de padrões no texto.	REM, resolução de anáforas
IDSAY	Classificação da pergunta em categorias.	Colecções de textos. Recuperação de documentos e ST tem por referência as entidades mencionadas e termos relevantes.	Por reconhecimento de entidades mencionadas e procura superficial de padrões no texto. Fusão de respostas próximas por concatenação do texto.	REM, ST, mecanismo de indexação pensado modularmente para isolar aspectos dependentes da língua.
PRIBERAM	Classificação da pergunta em categorias.	Colecções de textos. Textos indexados previamente e associação frase / (categoria + domínio conceptual) para a recuperação de texto.	Análise linguística intensa e pesquisa de padrões textuais ditados pela categoria de pergunta.	REM, ST, Ontologia de conceitos associada à fase de recuperação de texto, recurso FLiP.
SENSO	Classificação da pergunta em categorias. Classificador flexível.	Fontes: *Web*, ontologias, textos locais, dicionários e recursos especializados para algumas categorias. Recuperação de documentos: Google para a *Web*; sistema próprio, baseado no Lucene para os textos locais. Expansão da expressão de pesquisa.	Regras especializadas por categoria para procura superficial no texto, complementada com análise linguística localizada, e para inferência de respostas a partir de pistas (como datas relativas).	A lista resultados, ordenada por grau de confiança, é reapreciada por técnicas que favorecem respostas que se julgam mais correctas. Analisador PALAVRAS. Outros: REM, ST.

Tabela 5.1: Resumo de características do sistema SENSO e alguns SPR

Aqueles cujas fontes são apenas os documentos locais adoptam uma metodologia com maior ênfase na análise linguística, eventualmente no pré-processamento dos textos. É o caso do sistema PRIBERAM. No sistema SENSO, a análise linguística é aplicada de modo selectivo, sobre o texto envolvente às expressões detectadas como possíveis respostas pela procura superficial de padrões.
Relativamente à extracção, todos os sistemas da tabela incluem a procura de padrões no texto. Esta procura é guiada por regras, que podem ser mais simples nos sistemas PERGUNTE e XISQUÊ, ou mais elaboradas nos restantes sistemas. Por exemplo o sistema PRIBERAM aplica estas regras a frases de acordo com a categoria de pergunta e que tenham sido também associadas a um domínio conceptual, proveniente de uma ontologia. O sistema SENSO segue ainda outra abordagem na fase de extracção. Recorrendo à sua base de conhecimento inicial, a informações semânticas captadas em ontologias, e a elementos de informação em textos locais ou da *Web*, o sistema tenta responder com base num processo de inferência que conjuga os vários elementos disponíveis.

Com a excepção dos dois primeiros SPR da tabela 5.1, todos os outros participaram nas edições recentes do evento de pergunta-resposta QA@CLEF, para o Português. O QA@CLEF permitiu a comparação da eficácia na metodologia de cada sistema, com a avaliação dos respectivos resultados para o mesmo conjunto de perguntas. Apesar da organização do evento dar vários dias para o processamento das perguntas, o que flexibiliza a participação e dissipa a preocupação com a eficiência, no que respeita ao tempo necessário para as respostas, o QA@CLEF permite avaliar as respostas encontradas numa vasta colecção de textos, que é igual em todos os sistemas.
Nos anos 2007 e 2008, o sistema PRIBERAM alcançou o maior número de respostas correctas. Em ambas as edições, o sistema SENSO registou o segundo melhor grau de acerto global. E em 2008 o sistema associado a esta dissertação conseguiu o maior número de acertos para o tipo *Definição* [FPA+08].

Apesar do sistema SENSO não usar explicitamente a técnica de projecção de respostas, mencionada na secção 2.2 e usada por exemplo pelos sistemas QUALIM e PRISM, a metodologia seguida tem um efeito aproximado aquela técnica. Em primeiro lugar, a extracção de respostas tira partido da redundância existente nos documentos da *Web*, independentemente de outros processos. Depois, a valorização de respostas para as quais é detectada uma confirmação noutras fontes, possivelmente documentos locais, dicionários, ou repositórios especializados, contribui para eliminar as falsas respostas. O processo pode ser encarado como uma projecção de respostas não focada, onde a origem e o destino da projecção não são rígidos.
Consideremos agora a apreciação de respostas candidatas. A metodologia de [Lin02] foi pensada apenas para o tipo *Definição* e é baseada numa consulta directa no recurso WordNet e num processo estatístico sobre a ocorrência de termos. A técnica do sistema SENSO para as definições que é baseada na frequência de termos tem alguma semelhança com a segunda parte daquela técnica. Mas naquele sistema os termos contabilizados são de cada resposta face ao encontrado por um motor de busca na *Web*, e não entre

as próprias respostas candidatas. Já as técnicas principais deste modelo são focadas na semântica da resposta, sendo aplicáveis a qualquer categoria de pergunta.
O sistema QAAM faz uso de um grafo onde os elos representam relações de *equivalência* ou *inclusão*, enquanto que o sistema associado a esta dissertação prevê um conjunto de relações mais vasto. Outra diferença reside nas respostas, que no sistema QAAM são expressas por uma partição do grafo. O sistema SENSO dedica um nó a cada resposta candidata, que possui depois um conjunto de relações semânticas com outros nós do espaço. Aquele sistema aplica a referida metodologia apenas a locais e definições.

5.3 Balanço

A metodologia proposta no capítulo 3 para o funcionamento de um SPR, em geral, e as técnicas de selecção de respostas, em particular, foram avaliadas como se refere no capítulo 4. É altura de fazer um balanço deste trabalho, considerando os resultados obtidos.

5.3.1 Apreciação dos resultados do sistema

Efectuando uma apreciação geral, a participação no QA@CLEF, referida na secção 4.1.3, revela que o sistema SENSO consegue um bom nível de correcção nas suas respostas, tendo em conta os resultados de sistemas semelhantes sobre as mesmas questões.
A percentagem de acerto por categoria é superior para o tipo *Definição* e depois para as respostas às questões da classe *Quantidade*. Isto deve-se à natureza dessas respostas, pela relativa simplicidade na identificação das mesmas no conteúdo das fontes. Para as restantes categorias, as regras extracção detectam mais valores potencialmente interessantes, mas muitas vezes errados, o que diminui a a taxa de acerto final.

As técnicas propostas para a reapreciação de respostas candidatas melhoram os resultados do sistema, como se pode observar na secção 4.2. As técnicas baseadas na afinidade semântica levam a ganhos de aproximadamente 20% no acerto da melhor resposta do sistema, para as categorias *QueQual* e *Definição*, e um pouco menos para a categoria *Localização*. Nessas técnicas, e para as categorias *QueQual* e *Definição*, nota-se que a variante de semelhança de espectro com limite de bonificação 15 leva a ganhos significativos em ambos os casos. Já as variantes de bonificação com base no espectro combinado têm um efeito diferente nestas categorias. Em *QueQual*, a taxa de acerto não melhora ou chega mesmo a piorar. Mas em contrapartida, para a categoria *Definição* essa técnica leva a uma acentuada melhoria. Analisando os conjuntos de respostas candidatas para perguntas de ambas as categorias, constata-se um facto que justifica esta diferença no efeito da aplicação daquela técnica em particular. As respostas extraídas para questões do tipo *Definição* formam um conjunto geralmente mais homogéneo, do ponto de vista semântico, que o conjunto de respostas para *QueQual*. Como as respostas desta última categoria tendem a ser mais dispersas, a técnica de comparação resposta a

resposta funciona bem, mas o espectro combinado é mais indefinido. A relativa homogeneidade para as definições permite que as tendências do conjunto de resultados apontem para direcções correctas, o que se reflecte no efeito positivo da técnica para esses casos. Continuando nestas duas categorias, que entre as usadas para a avaliação das técnicas são as representativas das respostas textuais em geral, a técnica de confirmação por pesquisa na *Web* contribui positivamente para aumentar a taxa de acerto do sistema, em ambos os casos. Mas o ganho é mais acentuado no tipo *QueQual*, com uma subida de 20%, talvez porque a maioria daquelas respostas corresponde a factóides, com uma estrutura mais propícia para formar uma expressão de pesquisa.

Relativamente à categoria *Localização*, ambas as técnicas de bonificação de respostas com base na propagação de activação ao longo da malha semântica melhoram os resultados da fase de extracção. Os cinco indicadores na tabela 4.12 mostram que a técnica de semelhança de espectro de activação, na variante com limite de bonificação 15, é mais eficaz que qualquer variante sobre o espectro combinado. Como se observou para a categoria *QueQual*, o conjunto de respostas para este género de questão parece não evidenciar tendências globais que permitam tirar partido do espectro de activação combinado. A afinidade que possa existir entre algumas respostas deste género manifesta-se localmente, entre pares ou pequenos grupos, o que se reflecte na comparação dos respectivos espectros de activação. A confirmação por pesquisa na *Web* também revelou um contributo positivo para o acerto do sistema nesta categoria, mas aplicada em acumulação com as técnicas anteriores não supera o resultado que cada uma daquelas proporciona individualmente.

A aplicação da técnica baseada na afinidade semântica sobre as respostas candidatas na categoria *Quantidade* levou a um ganho reduzido, melhorando o acerto em aproximadamente 6%. A categoria *Quando* registou o menor benefício com a mesma técnica, cerca de 5% de incremento no acerto da primeira resposta de cada pergunta. A combinação da técnica de afinidade semântica com a confirmação de resposta por pesquisa na *Web* eleva essa melhoria até cerca de 10%. Para a categoria *Quantidade* não se registaram melhorias relevantes com o processo de confirmação por pesquisa na *Web*. Esta diferença pode ter a ver com o facto de existirem na *Web* bastantes documentos com dados biográficos, efemérides ou cronologias que repetem datas de eventos ou de acontecimentos populares, como muitos dos referidos na colecção de perguntas usada para os testes da categoria *Quando*. Mas não há elementos que permitam confirmar efectivamente esta hipótese.

As técnicas de apreciação de resultados com base na forma de cada resposta candidata também se revelaram úteis na melhoria dos resultados da categoria *Definição*. Concretamente, a terceira variante da técnica baseada na frequência de termos maximiza o acerto do sistema para aquela categoria, incrementando esse indicador em cerca de 20%. Duas variantes da técnica que verifica a presença de nomes no texto das definições também resultam num aumento considerável do acerto, de 14% e de 18%, como mostra a ta-

bela 4.8. Em ambas as técnicas destaca-se ainda um aumento acima de 10% no número de respostas candidatas correctas entre os cinco primeiros resultados para cada pergunta.

Apesar dos resultados serem bons face ao conseguido por outros sistemas semelhantes, a verdade é que continuam abaixo do que seria ideal, que é responder acertadamente a todas as questões de estrutura frásica simples, que não tenham ironia, anáforas ou outros aspectos que potenciem o equívoco. Num sistema complexo como é o caso de um SPR, os problemas podem manifestar-se de diversas maneiras e ao longo de várias etapas, sendo que uma falha na fase X irá repercutir-se ao longo das fases seguintes. Em seguida enumeram-se algumas das dificuldades detectadas durante a avaliação do sistema SENSO:

- **classificação**
  Algumas variantes menos comuns de perguntas ficam incorrectamente associadas à classe *TipoPorOmissão* quando deveriam ser classificadas numa das categorias mais concretas da secção 3.1.5. O classificador analisa a estrutura morfo-sintáctica da questão em busca das configurações mais comuns, havendo muito espaço para aperfeiçoamento.
  Outra falha ao nível da classificação tem a ver com as características identificadas na categoria que se atribui à pergunta. Por exemplo, o foco da questão pode não ser correctamente identificado ou as unidades pedidas podem não ser captadas numa pergunta do tipo *Quantidade*. Nesse caso a extracção de respostas não irá prosseguir da melhor forma.

- **recuperação**
  A expansão da expressão de pesquisa para a recuperação de documentos, via motor de busca na *Web* ou através do motor de pesquisa local, depende da informação semântica disponível para a versão de base da consulta. Questões que usam termos num segundo sentido, diferente do significado literal, podem ser afectadas pela expansão dos termos de pesquisa, a menos que algum dos recursos consultados pelo sistema permita a interpretação pretendida.

- **extracção**
  A extracção de respostas é controlada por regras, próprias de cada categoria de pergunta, que procuram elementos de informação que observem determinadas condições face à questão, e captam um resultado que provavelmente seria também escolhido por um leitor humano. Este processo manifesta dois tipos de falha. Por um lado, escapam algumas respostas cujo enquadramento nas fontes de informação não foi ainda considerado e associado a uma regra de extracção própria. Depois, para as regras existentes, há o inconveniente da extracção de falsas respostas. É difícil resolver este problema sem causar outro, que é a inibição da captura de respostas válidas. O aumento do grau de análise linguística poderá ajudar, mas

terá também impacto na eficiência.

- **inferência**
  As regras que utilizam inferência para encontrar a resposta a partir de evidências requerem um conjunto de elementos de informação que nem sempre se conseguem encontrar. Para evitar que o sistema fique indeterminadamente em ciclo na busca desses elementos, o processo é restringido a um pequeno número de tentativas, por sua vez limitadas no número de elementos de informação a considerar. Daqui resulta que algumas respostas poderiam ser encontradas considerando apenas mais um facto, ou com igual ou menor número de elementos de informação, mas simplesmente na tentativa que se seguiria.

- **reapreciação dos resultados**

  – contexto temporal e espacial
    A utilização dos contextos temporal e espacial na escolha ou reordenação do conjunto de respostas é exequível. A dificuldade reside na determinação destes contextos. Para a generalidade dos documentos utilizados, esta implementação do sistema não consegue apurar esses valores. Como a análise dos textos obtidos da *Web* é realizada na hora, notaram-se casos em que uma análise ideal do texto permitiria apurar o contexto temporal (ou espacial) sobre um elemento de informação, mas a indicação temporal que enquadra uma asserção no eixo do tempo ficou separada do trecho de texto seleccionado para a fase de extracção.
    Duas fontes onde esta contextualização foi aplicada com relativo sucesso foram duas colecções com textos jornalísticos, de um jornal português e outro brasileiro, usando os respectivos países como referência espacial e a data da publicação como referência para o contexto temporal.

  – contexto semântico
    A representação das respostas no eixo semântico do espaço de resultados pode ser afectada por alguns erros. No primeiro nível de expansão da malha de resultados para os factóides, a identificação dos termos de núcleo na definição do termo nem sempre é correcta. O mesmo sucede relativamente às respostas longas como as definições.
    Depois, as relações importadas desde um nó resposta, para formar a malha semântica, podem incluir alguns erros, o que interfere pontualmente com a apreciação dessa resposta e afecta indirectamente outras respostas que deixam de estar (ou passam a estar) com ela relacionadas. Detectaram-se também situações de importação de nós com excesso de relações com outros termos de reduzido valor informativo, mas que pela força do número das relações entre eles estabelecidas acabavam por beneficiar indevidamente com a propagação do nível de activação.

- valorização das parecenças na forma
  A utilização de uma técnica de bonificação de respostas candidatas com base na semelhança da forma ou constituição requer alguns cuidados. Tais técnicas aplicam-se por exemplo a resultados de definições. Alguns resultados extraídos incorrectamente, em particular de documentos *Web*, incluiam, a meio ou no fim, alguns símbolos ou expressões comuns ao tipo de documento de origem. Esses elementos aumentavam a correlação entre as respostas, o que prejudivava o resultado final do sistema por favorecer aquelas respostas erradas. À medida que foram detectados, estes casos resolveram-se ignorando as expressões referidas na comparação entre as respostas candidatas. Do mesmo modo, as *stop words* não são consideradas nesta técnica.
- valorização da afinidade semântica
  Tal como a popularidade não é sinónimo de correcção ou credibilidade, o uso da afinidade semântica para promover respostas, tendendo para resultados associados a um significado de maioria, pressupõe que essa maioria esteja correcta. Perguntas com resultados errados e semanticamente próximos podem sair prejudicadas na aplicação deste princípio. Para minimizar este efeito, a fase de extracção deve tão eficaz quanto possível.

5.3.2 Contribuições

Independentemente dos problemas pontuais referidos na secção anterior, a análise experimental revelou a utilidade da metodologia proposta. Numa perspectiva mais abrangente sobre este trabalho de investigação, consideramos que as principais contribuições são:

- **caracterização de técnicas, metodologias e sistemas**
  Descreveram-se os componentes existentes numa arquitectura típica de SPR: análise da pergunta, recuperação de documentos, análise do respectivo conteúdo e extracção de respostas. Foi elaborado um levantamento de técnicas aplicadas em cada uma destas fases, enquadradas em diferentes abordagens, com a indicação de alguns sistemas em que são empregues.

- **modelo teórico com inspiração humana**
  Responder a uma pergunta, num sistema de âmbito geral, envolve normalmente o processamento de documentos escritos em língua natural. Não há melhor entendedor da linguagem humana que o próprio Homem. Pelo senso comum, se uma pessoa não tem a resposta a uma pergunta, então documenta-se, por exemplo numa enciclopédia, ou pesquisando na *Web*. Esse mesmo procedimento pode ser usado, não para encontrar a resposta, mas para obter uma confirmação em caso de incerteza. Perante várias possíveis respostas, é consensual que uma pessoa faz uma ponderação sobre a validade de cada uma. A tendência é escolher a resposta que se mostrar mais apropriada, em função das informações que lhe estão associadas na sua memória, incluindo os significados possíveis e cenários ou contextos

de incidência do respectivo conceito. O modelo tenta imitar a referência humana, adoptando procedimentos correspondentes a:

- procura de respostas quando a memória não tem a solução
- confirmação pela consulta de fontes de informação, incluindo a *Web*
- determinação da resposta por analogia entre os resultados possíveis, com base nas informações disponíveis, que são em primeiro lugar as recordações suscitadas na memória por cada um desses resultados.

- **contextualização plena**
  As respostas candidatas são caracterizadas de uma forma mais rica que o usual, sendo associadas não só a um contexto semântico, mas também a contextos temporal e espacial. A definição deste ambiente torna mais natural a aceitação de diferentes respostas correctas para a mesma pergunta, cada uma válida em determinada região geográfica, por exemplo. O contexto temporal dota o modelo da expressividade para considerar a evolução das respostas ao longo do tempo.

- **espaço multidimensional de respostas**
  Para dar corpo ao contexto em que se inserem os resultados, o espaço multidimensional de respostas é a estrutura concreta que armazena respostas candidatas e permite a comparação das mesmas sob diversas perspectivas. Ao facilitar a diferenciação entre os resultados da fase de extracção, a vários níveis, esta estrutura serve de suporte às técnicas de apreciação de respostas candidatas, com o objectivo de melhorar a taxa de acerto do sistema.

- **rede semântica e propagação de activação**
  Este modelo recorre a uma rede semântica para representação de respostas candidatas, em particular para o eixo semântico do espaço multidimensional e para as respostas formadas por texto. Com esta estrutura associativa, é possível agregar todos os resultados a considerar e ainda elementos semânticos individuais ou comuns a várias respostas. A malha semântica, resultante do processo de expansão dos nós resposta, permite lidar com a ambiguidade através da representação de todos os significados conhecidos para um conceito. Nessas situações, o significado relevante encontrará eco noutra resposta candidata, eventualmente, emergindo entre os demais. A propagação de activação é o procedimento prático para a identificação destes significados e outros elementos semânticos relevantes para cada resposta candidata, necessários para a reapreciação dos resultados da fase de extracção.

- **inferência como complemento à extracção directa**
  O modelo responde às perguntas não apenas pelo fornecimento de resultados encontrados em fontes de informação, qualquer que seja a natureza das mesmas, mas

também com respostas calculadas, resultantes de um processo de inferência. O processo de resposta complementa a extracção directa com regras lógicas que combinam múltiplos elementos de informação, no sentido de inferir uma solução para a pergunta. A base de conhecimento inicial do sistema é um repositório semântico essencial ao motor de inferência.

- **proposta e avaliação de técnicas de apreciação de respostas candidatas**
  Para além dos princípios gerais do modelo, foram descritas técnicas concretas para a apreciação de respostas candidatas num SPR. Tirando partido da organização que o modelo prevê para os resultados provisórios, estas técnicas são de uso geral para respostas de texto. Quantidades e datas podem ser processadas com o mesmo princípio, mudando apenas a implementação. Estas técnicas foram ainda avaliadas para um conjunto representativo de perguntas de várias categorias, com o resultado apresentado no capítulo 4.

- **sistema** SENSO
  Para além do modelo, dos princípios que influenciam a respectiva metodologia e das técnicas delineadas para resolver aspectos chave no tratamento de uma pergunta, este trabalho de investigação resultou também na implementação do sistema SENSO. Mais que um protótipo para testar técnicas específicas, o sistema desenvolvido é efectivamente um SPR. Realçamos ainda o facto deste sistema ter sido aplicado ao Português, não por ser um requisito, mas por opção. A metodologia, em termos gerais, não depende da língua. Isso sucede apenas em elementos concretos, como os padrões textuais usados nas regras de análise textual ou o analisador morfo-sintáctico, mas poderiam adaptar-se para outros idiomas.

5.4 Trabalho Futuro

Num trabalho cujo modelo tenta imitar o comportamento inteligente dos humanos na escolha de respostas acertadas, é natural que a automatização dessa faculdade, própria das pessoas, fique aquém do ideal, deixando espaço para diversas linhas de desenvolvimento do modelo. Também no plano prático, o extenso e complexo processo de responder automaticamente a perguntas, de forma concisa, é composto por algumas tarefas específicas cuja implementação pode e deve ser aperfeiçoada, designadamente para resolver algumas dificuldades mencionadas na secção 5.3.1.

O algoritmo de propagação do nível de activação semântica, aplicado sobre a malha com o espaço de resultados, pode ser alvo de um estudo que conduza à sua actualização, para se obter um resultado mais correcto. Um dos aspectos chave é a ponderação que se usa quando um nó é alvo de múltiplas activações. Em particular, a escolha entre uma versão cumulativa, onde o nível de activação cresce proporcionalmente ao número de

activações, e no outro extremo, a adopção de um valor médio de activação. A primeira versão tem um efeito por vezes indesejado, que é a activação excessiva de nós pouco relevantes, pelo simples facto de terem muitas relações. A utilização da média dos valores de propagação também tem o inconveniente de uma activação de baixo valor contrariar a intenção de aumento do nível de activação para aquele nó. Para além do estudo sobre as variantes do algoritmo de propagação, também os coeficientes de propagação do nível de activação, específicos de cada relação semântica e indicados na tabela 3.1, carecem de um estudo que sustente a escolha dos respectivos valores.

As respostas alusivas a tempo, horas ou época podem ter um tratamento semântico para além da apreciação quantitativa da distância entre datas, por exemplo. Expressões como "*à hora de jantar*", "*ao anoitecer*" ou "*no fim do dia*" estão implicitamente relacionadas. Isto pode e deve ser obtido pelo sistema na base de conhecimento, mais concretamente pela relação *relacionadoCom*. Mas para além disso, aquelas expressões podem ainda ser relacionadas com uma hora que pelo senso comum associamos aqueles conceitos, que não tem de ser um só valor específico mas pode situar-se num intervalo como o período 19:00 - 21:00. Todas estas associações semânticas que se desencadeiam na nossa mente deviam ser consideradas na activação semântica sobre o espaço de resultados. Para tal, é necessário um bom repositório semântico, composto de factos e também de regras, directamente na base de conhecimento, que facilitem a inferência.

A pergunta poderia ser também considerada na malha que representa o eixo semântico do espaço de resultados. Não é claro que potenciais associações entre nós resposta e elementos na pergunta sejam pistas para as melhores respostas. Muitas vezes os termos empregues na questão não têm relação directa com a resposta. Ainda assim, seria interessante explorar este caminho.

A bonificação de respostas em função da afinidade semântica no sistema SENSO não depende dos contextos temporal e espacial. Esses são verificados separadamente. Valeria a pena definir vários critérios para a aplicação da bonificação por proximidade semântica, por exemplo em função também da concordância espacial e temporal das respostas e efectuar uma avaliação de larga escala ao uso desses critérios. A intenção seria considerar a semelhança de significados apenas se as respostas são válidas na mesma época e/ou zona geográfica.

A combinação de técnicas de apreciação de respostas conduziu a resultados menos interessantes que os obtidos por uma técnica apenas, para alguns casos. Era útil dispor de uma forma de combinar técnicas diferentes que evitasse este cancelamento recíproco das melhorias relativas e aumentasse o acerto global.

O modelo podia permitir a interacção com o utilizador, na fase final do processo, após a apresentação das melhores respostas do sistema. Se o utilizador indicar qual a resposta ou respostas que considera correctas, o sistema pode ajustar dinamicamente a

sua base de conhecimento. A operação de cálculo da nova graduação para as respostas pode mudar substancialmente, em função de elementos recolhidos nas indicações do utilizador. Para além de assegurar o acerto em futuros processamentos sobre a mesma questão, esta evolução nas crenças do sistema poderá melhorar também os resultados de perguntas semelhantes ou sobre o mesmo domínio semântico.
Porque utilizadores diferentes podem ter preferências distintas e opiniões eventualmente contraditórias, o sistema deve separar o contexto de cada utilizador. No limite, esta funcionalidade é como uma personalização do processo de escolha de respostas.

Outro aspecto interessante é a optimização de um SPR, desta vez não no sentido da eficácia, mas antes da eficiência, com o intuito de diminuir o tempo de resposta. A demora em obter a resposta num SPR continua a ser um factor de desânimo para o utilizador comum. Esta é uma das razões para alguns sistemas *online* terem uma solução de último recurso, apresentada ao fim de um tempo limite, ou ainda quando não encontram uma resposta, que não é um resultado conciso mas antes um bloco de texto obtido num motor de busca. Num sistema como o SENSO, que faz uso de fontes de informação locais e através da rede, e cuja metodologia prevê procedimentos multi-abordagem, seria útil realizar um estudo que refizesse a arquitectura, identificando os procedimentos que se podem executar em paralelo. O recurso a serviços auxiliares externos ao SPR tem impacto no tempo de resposta, pelo que deve ser pensado de raiz e optimizado através do uso de *cache*[1]. O sistema SENSO usa *cache*, nomeadamente nas operações frequentes durante o processamento de uma pergunta, como a consulta de definições em dicionários e o teste de compatibilidade semântica entre dois termos.

Em suma, a avaliação da metodologia proposta para o funcionamento de um SPR mostrou que o trabalho apresentado nesta dissertação é um contributo para a melhoria dos resultados dos sistemas. Para lá dos melhoramentos possíveis na implementação, este modelo teórico pode evoluir em várias direcções, nomeadamente através de estudos mais aprofundados sobre a activação semântica aplicada à selecção de respostas candidatas ou pela introdução de novas técnicas de apreciação de resultados, inspiradas ou não no comportamento humano.
O trabalho desenvolvido abrange vários domínios, nomeadamente linguística, informática e as neurociências, cuja combinação no âmbito dos SPR é complexa mas ao mesmo tempo aliciante. Podemos considerar que os objectivos estabelecidos foram alcançados. Embora exista um longo caminho a percorrer, os SPR de domínio aberto tiram cada vez mais partido das fontes de informação e aproximam-se cada vez mais do pretendido pelo utilizador.

[1] *Cache*: réplica de dados necessários ao funcionamento de um sistema, usada para aumentar a eficiência. Quando é possível, a existência de dados suficientes na réplica evita a reexecução das operações que conduziriam a esses mesmos dados, o que contribui para reduzir o tempo de processamento.

Apêndice A

Outros resultados de Avaliação

Este anexo contém uma lista de tabelas com resultados detalhados da avaliação às técnicas de reapreciação das respostas candidatas. Estas tabelas complementam os resultados apresentados na secção 4.2.2.

Categoria QueQual

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	sim	1	2	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Qual é a montanha mais alta do México?	não	6	1	(0/5) 0,00	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	6	1	(0/5) 0,00	(1/10) 0,10
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	não	3	1	(1/5) 0,20	(1/10) 0,10
Que rio nasce na Serra da Estrela?	sim	1	2	(1/5) 0,20	(6/10) 0,60
Que prémio recebeu José Saramago?	sim	1	3	(4/5) 0,80	(8/10) 0,80
Que tipo de ave é o milhafre-preto?	não	2	1	(3/5) 0,60	(6/10) 0,60
média	73,33%	1,87	2,00	30,67%	30,44%

Tabela A.1: QueQual: semelhança de espectro com bonificação crescente

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(3/5) 0,60	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	11	1	(0/5) 0,00	(0/10) 0,00
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	não	5	1	(1/5) 0,20	(1/10) 0,10
Que rio nasce na Serra da Estrela?	sim	1	2	(1/5) 0,20	(6/10) 0,60
Que prémio recebeu José Saramago?	sim	1	3	(2/5) 0,40	(6/10) 0,60
Que tipo de ave é o milhafre-preto?	sim	1	2	(4/5) 0,80	(7/10) 0,70
média	80,00%	2,00	2,07	30,67%	29,11%

Tabela A.2: QueQual: semelhança de espectro com bonificação limitada a 15

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	9	1	(0/5) 0,00	(1/10) 0,10
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	não	2	1	(1/5) 0,20	(1/10) 0,10
Que rio nasce na Serra da Estrela?	sim	1	2	(1/5) 0,20	(6/10) 0,60
Que prémio recebeu José Saramago?	sim	1	3	(2/5) 0,40	(6/10) 0,60
Que tipo de ave é o milhafre-preto?	sim	1	2	(4/5) 0,80	(6/10) 0,60
média	80,00%	1,67	2,07	29,33%	29,11%

Tabela A.3: QueQual: semelhança de espectro com bonificação limitada a 30

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	9	1	(0/5) 0,00	(1/10) 0,10
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	não	2	1	(1/5) 0,20	(1/10) 0,10
Que rio nasce na Serra da Estrela?	sim	1	2	(1/5) 0,20	(6/10) 0,60
Que prémio recebeu José Saramago?	sim	1	3	(4/5) 0,80	(6/10) 0,60
Que tipo de ave é o milhafre-preto?	sim	1	2	(4/5) 0,80	(6/10) 0,60
média	80,00%	1,67	2,07	32,00%	29,11%

Tabela A.4: QueQual: semelhança de espectro com bonificação limitada a 45

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	6	1	(0/5) 0,00	(1/10) 0,10
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	não	3	1	(1/5) 0,20	(1/10) 0,10
Que rio nasce na Serra da Estrela?	sim	1	2	(1/5) 0,20	(6/10) 0,60
Que prémio recebeu José Saramago?	não	2	1	(4/5) 0,80	(8/10) 0,80
Que tipo de ave é o milhafre-preto?	sim	1	2	(4/5) 0,80	(6/10) 0,60
média	73,33%	1,60	1,93	32,00%	30,44%

Tabela A.5: QueQual: semelhança de espectro com bonificação limitada a 100

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	não	2	1	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	não	4	1	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	não	4	1	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Qual é a língua oficial do Egipto?	não	2	1	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	não	2	1	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	não	3	1	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	23	1	(0/5) 0,00	(0/10) 0,00
Qual é a maior cidade do Canadá?	não	15	1	(0/5) 0,00	(0/10) 0,00
Que tipo de animal é o Cocas?	não	15	1	(0/5) 0,00	(0/10) 0,00
Que rio nasce na Serra da Estrela?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Que prémio recebeu José Saramago?	sim	1	4	(3/5) 0,60	(7/10) 0,70
Que tipo de ave é o milhafre-preto?	não	4	1	(2/5) 0,40	(5/10) 0,50
média	26,67%	5,33	1,40	26,67%	25,78%

Tabela A.6: QueQual: proximidade aos nós mais activos do espectro combinado

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	2	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	não	3	1	(2/5) 0,40	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	20	1	(0/5) 0,00	(0/10) 0,00
Qual é a maior cidade do Canadá?	não	10	1	(0/5) 0,00	(1/10) 0,10
Que tipo de animal é o Cocas?	não	8	1	(0/5) 0,00	(1/10) 0,10
Que rio nasce na Serra da Estrela?	não	3	1	(2/5) 0,40	(4/10) 0,40
Que prémio recebeu José Saramago?	sim	1	3	(2/5) 0,40	(6/10) 0,60
Que tipo de ave é o milhafre-preto?	sim	1	3	(4/5) 0,80	(7/10) 0,70
média	60,00%	3,67	1,87	29,33%	27,78%

Tabela A.7: QueQual: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 15

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	2	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	não	2	1	(2/5) 0,40	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	22	1	(0/5) 0,00	(0/10) 0,00
Qual é a maior cidade do Canadá?	não	12	1	(0/5) 0,00	(0/10) 0,00
Que tipo de animal é o Cocas?	não	10	1	(0/5) 0,00	(1/10) 0,10
Que rio nasce na Serra da Estrela?	não	2	1	(1/5) 0,20	(4/10) 0,40
Que prémio recebeu José Saramago?	sim	1	3	(2/5) 0,40	(6/10) 0,60
Que tipo de ave é o milhafre-preto?	sim	1	4	(4/5) 0,80	(6/10) 0,60
média	60,00%	3,93	1,93	28,00%	26,44%

Tabela A.8: QueQual: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 30

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(2/5) 0,40	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	2	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	3	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	não	4	1	(1/5) 0,20	(2/10) 0,20
Qual é a montanha mais alta do México?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio banha Paris?	não	2	1	(2/5) 0,40	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	23	1	(0/5) 0,00	(0/10) 0,00
Qual é a maior cidade do Canadá?	não	13	1	(0/5) 0,00	(0/10) 0,00
Que tipo de animal é o Cocas?	não	13	1	(0/5) 0,00	(0/10) 0,00
Que rio nasce na Serra da Estrela?	não	2	1	(1/5) 0,20	(4/10) 0,40
Que prémio recebeu José Saramago?	sim	1	3	(2/5) 0,40	(6/10) 0,60
Que tipo de ave é o milhafre-preto?	sim	1	4	(3/5) 0,60	(6/10) 0,60
média	53,33%	4,47	1,80	26,67%	25,78%

Tabela A.9: QueQual: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 45

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que constelação fica Vega?	não	2	1	(3/5) 0,60	(3/10) 0,30
Que pássaro renascia das próprias cinzas?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Que espada usavam as legiões romanas?	sim	1	4	(3/5) 0,60	(3/5) 0,60
Que ordem foi fundada por Santa Clara?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Qual é a montanha mais alta do México?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Que rio banha Paris?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Qual é a língua oficial do Egipto?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que guerra combateu Joana de Arc?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Qual a nacionalidade de Nicole Kidman?	sim	1	2	(1/5) 0,20	(1/6) 0,17
Que navio americano foi afundado em Pearl Harbor em 1941?	não	40	1	(0/5) 0,00	(0/10) 0,00
Qual é a maior cidade do Canadá?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que tipo de animal é o Cocas?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que rio nasce na Serra da Estrela?	sim	1	3	(3/5) 0,60	(5/10) 0,50
Que prémio recebeu José Saramago?	sim	1	5	(4/5) 0,80	(7/10) 0,70
Que tipo de ave é o milhafre-preto?	não	2	1	(3/5) 0,60	(7/10) 0,70
média	80,00%	3,73	2,33	37,33%	29,11%

Tabela A.10: QueQual: confirmação por pesquisa *Web*

Categoria Definição

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(3/5) 0,60	(8/10) 0,80
O que era o A6M Zero?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que é a paella?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	sim	1	2	(2/5) 0,40	(6/10) 0,60
O que é o feta?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é o jagertee?	não	3	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que era a RSFSR?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é um berimbau?	sim	1	4	(3/5) 0,60	(7/10) 0,70
O que é VRML?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que são os forcados?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	não	5	1	(1/5) 0,20	(2/10) 0,20
O que é que é um brigadeiro?	não	2	1	(4/5) 0,80	(8/10) 0,80
O que é a Feplam?	não	2	1	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é o Crescente Fértil?	não	2	1	(2/5) 0,40	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	6	1	(0/5) 0,00	(1/10) 0,10
O que são os iaques?	sim	1	4	(4/5) 0,80	(8/10) 0,80
média	72,73%	1,64	3,05	60,00%	50,86%

Tabela A.11: Definição: semelhança de espectro com bonificação crescente

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(3/5) 0,60	(8/10) 0,80
O que era o A6M Zero?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é a paella?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	sim	1	3	(2/5) 0,40	(6/10) 0,60
O que é o feta?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	não	2	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que era a RSFSR?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é um berimbau?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é VRML?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que são os forcados?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	2	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	6	1	(0/5) 0,00	(1/10) 0,10
O que são os iaques?	sim	1	8	(5/5) 1,00	(8/10) 0,80
média	86,36%	1,32	3,36	60,91%	49,95%

Tabela A.12: Definição: semelhança de espectro com bonificação limitada a 15

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(3/5) 0,60	(8/10) 0,80
O que era o A6M Zero?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é a paella?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	sim	1	3	(3/5) 0,60	(7/10) 0,70
O que é o feta?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é o jagertee?	não	3	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que era a RSFSR?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é um berimbau?	sim	1	3	(3/5) 0,60	(7/10) 0,70
O que é VRML?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que são os forcados?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	não	5	1	(1/5) 0,20	(2/10) 0,20
O que é que é um brigadeiro?	não	2	1	(4/5) 0,80	(8/10) 0,80
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	6	1	(0/5) 0,00	(1/10) 0,10
O que são os iaques?	sim	1	4	(4/5) 0,80	(8/10) 0,80
média	81,82%	1,55	3,00	60,91%	51,31%

Tabela A.13: Definição: semelhança de espectro com bonificação limitada a 30

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	3	(4/5) 0,80	(5/10) 0,50
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que era o A6M Zero?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é a paella?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que é a açorda?	sim	1	2	(2/5) 0,40	(6/10) 0,60
O que é o feta?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que era a RSFSR?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é um berimbau?	não	3	1	(2/5) 0,40	(6/10) 0,60
O que é VRML?	sim	1	3	(2/5) 0,40	(6/10) 0,60
O que são os forcados?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	não	6	1	(0/5) 0,00	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	7	(5/5) 1,00	(6/10) 0,60
O que é o Crescente Fértil?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que é o Gil Vicente FC?	não	6	1	(0/5) 0,00	(2/10) 0,20
O que é a Torre do Tombo?	não	16	1	(0/5) 0,00	(0/10) 0,00
O que são os iaques?	sim	1	2	(3/5) 0,60	(7/10) 0,70
média	81,82%	2,23	2,86	55,45%	46,31%

Tabela A.14: Definição: proximidade aos nós mais activos do espectro combinado

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(4/5) 0,80	(5/10) 0,50
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que era o A6M Zero?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que é a paella?	sim	1	2	(3/5) 0,60	(6/10) 0,60
O que é a açorda?	sim	1	3	(3/5) 0,60	(7/10) 0,70
O que é o feta?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que era a RSFSR?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é um berimbau?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é VRML?	sim	1	7	(5/5) 1,00	(6/10) 0,60
O que são os forcados?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é o Crescente Fértil?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é o Gil Vicente FC?	não	6	1	(0/5) 0,00	(2/10) 0,20
O que é a Torre do Tombo?	não	8	1	(0/5) 0,00	(2/10) 0,20
O que são os iaques?	sim	1	4	(3/5) 0,60	(7/10) 0,70
média	90,91%	1,55	3,41	59,09%	48,59%

Tabela A.15: Definição: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 30

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	3	(4/5) 0,80	(5/10) 0,50
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que era o A6M Zero?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é a paella?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que é a açorda?	sim	1	2	(2/5) 0,40	(6/10) 0,60
O que é o feta?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que era a RSFSR?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é um berimbau?	não	4	1	(2/5) 0,40	(6/10) 0,60
O que é VRML?	sim	1	3	(2/5) 0,40	(6/10) 0,60
O que são os forcados?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	7	(5/5) 1,00	(6/10) 0,60
O que é o Crescente Fértil?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que é o Gil Vicente FC?	não	6	1	(0/5) 0,00	(2/10) 0,20
O que é a Torre do Tombo?	não	15	1	(0/5) 0,00	(0/10) 0,00
O que são os iaques?	sim	1	2	(3/5) 0,60	(7/10) 0,70
média	86,36%	2,00	2,95	57,27%	46,31%

Tabela A.16: Definição: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 45

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	não	7	1	(0/5) 0,00	(1/10) 0,10
O que é a Brabançonne?	sim	1	4	(4/5) 0,80	(4/10) 0,40
O que é uma cítara?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que era o A6M Zero?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é a paella?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	não	2	1	(2/5) 0,40	(7/10) 0,70
O que é o feta?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que é o IPM em Portugal?	sim	1	7	(5/5) 1,00	(6/10) 0,60
O que era a RSFSR?	sim	1	4	(4/5) 0,80	(7/10) 0,70
O que é um berimbau?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é VRML?	sim	1	2	(3/5) 0,60	(6/10) 0,60
O que são os forcados?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	não	41	1	(0/5) 0,00	(0/10) 0,00
O que é que é um brigadeiro?	não	2	1	(2/5) 0,40	(4/10) 0,40
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	5	(4/5) 0,80	(5/10) 0,50
O que é o Crescente Fértil?	sim	1	5	(4/5) 0,80	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	13	1	(0/5) 0,00	(0/10) 0,00
O que são os iaques?	sim	1	6	(5/5) 1,00	(7/10) 0,70
média	72,73%	3,77	3,36	59,09%	44,95%

Tabela A.17: Definição: frequência de termos 1

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	não	5	1	(1/5) 0,20	(2/10) 0,20
O que é a Brabançonne?	sim	1	4	(4/5) 0,80	(4/10) 0,40
O que é uma cítara?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que era o A6M Zero?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é a paella?	não	2	1	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	não	2	1	(4/5) 0,80	(7/10) 0,70
O que é o feta?	sim	1	9	(5/5) 1,00	(8/10) 0,80
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que era a RSFSR?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é um berimbau?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é VRML?	sim	1	11	(5/5) 1,00	(10/10) 1,00
O que são os forcados?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	não	35	1	(0/5) 0,00	(0/10) 0,00
O que é que é um brigadeiro?	não	3	1	(1/5) 0,20	(3/10) 0,30
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	6	(5/5) 1,00	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	5	(4/5) 0,80	(5/10) 0,50
O que é o Gil Vicente FC?	não	5	1	(1/5) 0,20	(1/10) 0,10
O que é a Torre do Tombo?	não	5	1	(1/5) 0,20	(2/10) 0,20
O que são os iaques?	sim	1	8	(5/5) 1,00	(7/10) 0,70
média	68,18%	3,27	3,86	63,64%	48,13%

Tabela A.18: Definição: frequência de termos 2: filtragem

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	não	2	1	(3/5) 0,60	(5/10) 0,50
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que era o A6M Zero?	sim	1	5	(4/5) 0,80	(8/10) 0,80
O que é a paella?	sim	1	4	(3/5) 0,60	(5/10) 0,50
O que é a açorda?	sim	1	3	(2/5) 0,40	(6/10) 0,60
O que é o feta?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é o IPM em Portugal?	sim	1	2	(3/5) 0,60	(4/10) 0,40
O que era a RSFSR?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que é um berimbau?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que é VRML?	não	2	1	(3/5) 0,60	(8/10) 0,80
O que são os forcados?	não	3	1	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	não	2	1	(1/5) 0,20	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é o Crescente Fértil?	sim	1	2	(1/5) 0,20	(4/10) 0,40
O que é o Gil Vicente FC?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	3	1	(1/5) 0,20	(1/10) 0,10
O que são os iaques?	sim	1	5	(4/5) 0,80	(6/10) 0,60
média	77,27%	1,32	2,86	55,45%	46,77%

Tabela A.19: Definição: confirmação por pesquisa *Web* 1

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	5	(4/5) 0,80	(4/10) 0,40
O que é uma cítara?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que era o A6M Zero?	sim	1	2	(2/5) 0,40	(7/10) 0,70
O que é a paella?	não	3	1	(2/5) 0,40	(3/10) 0,30
O que é a açorda?	sim	1	3	(2/5) 0,40	(7/10) 0,70
O que é o feta?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	5	(4/5) 0,80	(5/10) 0,50
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que era a RSFSR?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que é um berimbau?	sim	1	4	(3/5) 0,60	(7/10) 0,70
O que é VRML?	não	2	1	(3/5) 0,60	(7/10) 0,70
O que são os forcados?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	9	(5/5) 1,00	(8/10) 0,80
O que é a Feplam?	não	2	1	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	7	(5/5) 1,00	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	6	(5/5) 1,00	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	sim	1	2	(3/5) 0,60	(3/10) 0,30
O que são os iaques?	sim	1	6	(5/5) 1,00	(7/10) 0,70
média	81,82%	1,23	3,41	60,91%	48,13%

Tabela A.20: Definição: confirmação por pesquisa *Web* 2

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que era o A6M Zero?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é a paella?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que é o feta?	sim	1	4	(4/5) 0,80	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que era a RSFSR?	não	4	1	(1/5) 0,20	(4/10) 0,40
O que é um berimbau?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que é VRML?	sim	1	2	(4/5) 0,80	(7/10) 0,70
O que são os forcados?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	2	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	4	(3/5) 0,60	(7/10) 0,70
O que é o Crescente Fértil?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que é o Gil Vicente FC?	não	3	1	(1/5) 0,20	(2/10) 0,20
O que é a Torre do Tombo?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que são os iaques?	sim	1	6	(5/5) 1,00	(7/10) 0,70
média	90,91%	1,23	3,00	57,27%	47,22%

Tabela A.21: Definição: presença de nomes 1, minimo: 2

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	2	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	7	(5/5) 1,00	(8/10) 0,80
O que era o A6M Zero?	sim	1	7	(5/5) 1,00	(8/10) 0,80
O que é a paella?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	não	2	1	(3/5) 0,60	(5/10) 0,50
O que é o feta?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(3/10) 0,30
O que era a RSFSR?	não	2	1	(2/5) 0,40	(4/10) 0,40
O que é um berimbau?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que é VRML?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que são os forcados?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	4	(4/5) 0,80	(8/10) 0,80
O que é o Crescente Fértil?	não	2	1	(2/5) 0,40	(5/10) 0,50
O que é o Gil Vicente FC?	não	2	1	(1/5) 0,20	(2/10) 0,20
O que é a Torre do Tombo?	não	2	1	(3/5) 0,60	(4/10) 0,40
O que são os iaques?	sim	1	7	(5/5) 1,00	(7/10) 0,70
média	77,27%	1,23	3,32	62,73%	49,04%

Tabela A.22: Definição: presença de nomes 1, minimo: 4

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que era o A6M Zero?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é a paella?	não	4	1	(2/5) 0,40	(4/10) 0,40
O que é a açorda?	não	2	1	(2/5) 0,40	(5/10) 0,50
O que é o feta?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que era a RSFSR?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que é um berimbau?	sim	1	2	(2/5) 0,40	(4/10) 0,40
O que é VRML?	não	2	1	(3/5) 0,60	(5/10) 0,50
O que são os forcados?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	2	(3/5) 0,60	(7/10) 0,70
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	2	(3/5) 0,60	(6/10) 0,60
O que é o Crescente Fértil?	não	2	1	(1/5) 0,20	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(1/5) 0,20	(1/10) 0,10
O que é a Torre do Tombo?	não	4	1	(1/5) 0,20	(3/10) 0,30
O que são os iaques?	sim	1	5	(4/5) 0,80	(6/10) 0,60
média	72,73%	1,45	2,27	48,18%	42,22%

Tabela A.23: Definição: presença de nomes 2

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que era o A6M Zero?	sim	1	5	(4/5) 0,80	(8/10) 0,80
O que é a paella?	não	2	1	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	sim	1	3	(2/5) 0,40	(6/10) 0,60
O que é o feta?	sim	1	6	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(3/10) 0,30
O que era a RSFSR?	sim	1	2	(1/5) 0,20	(4/10) 0,40
O que é um berimbau?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é VRML?	não	2	1	(4/5) 0,80	(5/10) 0,50
O que são os forcados?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	2	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	4	(3/5) 0,60	(6/10) 0,60
O que é o Crescente Fértil?	não	3	1	(2/5) 0,40	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(1/5) 0,20	(2/10) 0,20
O que é a Torre do Tombo?	sim	1	2	(1/5) 0,20	(2/10) 0,20
O que são os iaques?	sim	1	6	(5/5) 1,00	(7/10) 0,70
média	86,36%	1,18	2,91	55,45%	45,86%

Tabela A.24: Definição: presença de nomes 3

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(3/5) 0,60	(7/10) 0,70
O que era o A6M Zero?	sim	1	2	(4/5) 0,80	(5/10) 0,50
O que é a paella?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é o feta?	sim	1	8	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	não	2	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que era a RSFSR?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é um berimbau?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é VRML?	sim	1	7	(5/5) 1,00	(8/10) 0,80
O que são os forcados?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(8/10) 0,80
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é o Gil Vicente FC?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	11	1	(0/5) 0,00	(0/10) 0,00
O que são os iaques?	sim	1	4	(4/5) 0,80	(7/10) 0,70
média	86,36%	1,55	3,27	62,73%	49,95%

Tabela A.25: Definição: semelhança de espectro + proximidade a nós mais activos do espectro combinado

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(4/5) 0,80	(4/10) 0,40
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	3	(3/5) 0,60	(7/10) 0,70
O que era o A6M Zero?	sim	1	5	(4/5) 0,80	(5/10) 0,50
O que é a paella?	sim	1	2	(4/5) 0,80	(6/10) 0,60
O que é a açorda?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é o feta?	sim	1	8	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(1/5) 0,20	(3/10) 0,30
O que era a RSFSR?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é um berimbau?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é VRML?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que são os forcados?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	sim	1	3	(3/5) 0,60	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é o Crescente Fértil?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é o Gil Vicente FC?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	6	1	(0/5) 0,00	(2/10) 0,20
O que são os iaques?	sim	1	4	(4/5) 0,80	(7/10) 0,70
média	90,91%	1,27	3,36	61,82%	48,13%

Tabela A.26: Definição: presença de nomes + proximidade a nós mais activos do espectro combinado

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	5	(4/5) 0,80	(4/10) 0,40
O que é uma cítara?	sim	1	2	(4/5) 0,80	(7/10) 0,70
O que era o A6M Zero?	não	2	1	(4/5) 0,80	(5/10) 0,50
O que é a paella?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é a açorda?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é o feta?	sim	1	8	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que era a RSFSR?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é um berimbau?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é VRML?	sim	1	6	(5/5) 1,00	(7/10) 0,70
O que são os forcados?	sim	1	2	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	6	(5/5) 1,00	(7/10) 0,70
O que é o Crescente Fértil?	sim	1	6	(5/5) 1,00	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(1/5) 0,20	(2/10) 0,20
O que é a Torre do Tombo?	não	3	1	(1/5) 0,20	(1/10) 0,10
O que são os iaques?	sim	1	3	(4/5) 0,80	(7/10) 0,70
média	90,91%	1,14	3,77	67,27%	49,04%

Tabela A.27: Definição: confirmação por pesquisa *Web* + proximidade a nós mais activos do espectro combinado

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	5	(4/5) 0,80	(4/10) 0,40
O que é uma cítara?	sim	1	3	(4/5) 0,80	(8/10) 0,80
O que era o A6M Zero?	não	2	1	(4/5) 0,80	(5/10) 0,50
O que é a paella?	sim	1	3	(3/5) 0,60	(5/10) 0,50
O que é a açorda?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é o feta?	sim	1	8	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	4	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(1/5) 0,20	(3/10) 0,30
O que era a RSFSR?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é um berimbau?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é VRML?	sim	1	6	(5/5) 1,00	(7/10) 0,70
O que são os forcados?	sim	1	2	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	sim	1	3	(3/5) 0,60	(3/10) 0,30
O que é que é um brigadeiro?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é o Crescente Fértil?	sim	1	6	(5/5) 1,00	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(1/5) 0,20	(2/10) 0,20
O que é a Torre do Tombo?	não	2	1	(1/5) 0,20	(3/10) 0,30
O que são os iaques?	sim	1	3	(4/5) 0,80	(8/10) 0,80
média	90,91%	1,09	3,77	66,36%	49,95%

Tabela A.28: Definição: confirmação por pesquisa *Web* + proximidade a nós mais activos do espectro combinado + presença de nomes

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	2	(3/5) 0,60	(8/10) 0,80
O que era o A6M Zero?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é a paella?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é a açorda?	sim	1	3	(2/5) 0,40	(6/10) 0,60
O que é o feta?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é o jagertee?	não	3	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que era a RSFSR?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é um berimbau?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é VRML?	sim	1	7	(5/5) 1,00	(8/10) 0,80
O que são os forcados?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	3	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	8	(5/5) 1,00	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	4	(3/5) 0,60	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	6	1	(0/5) 0,00	(2/10) 0,20
O que são os iaques?	sim	1	7	(5/5) 1,00	(8/10) 0,80
média	90,91%	1,32	3,77	63,64%	51,31%

Tabela A.29: Definição: frequência de termos + semelhança de espectro

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(4/5) 0,80	(4/10) 0,40
O que é a Brabançonne?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	4	(4/5) 0,80	(7/10) 0,70
O que era o A6M Zero?	sim	1	5	(4/5) 0,80	(5/10) 0,50
O que é a paella?	sim	1	6	(5/5) 1,00	(6/10) 0,60
O que é a açorda?	sim	1	3	(3/5) 0,60	(6/10) 0,60
O que é o feta?	sim	1	7	(5/5) 1,00	(8/10) 0,80
O que é o jagertee?	não	2	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(3/10) 0,30
O que era a RSFSR?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que é um berimbau?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é VRML?	sim	1	7	(5/5) 1,00	(6/10) 0,60
O que são os forcados?	sim	1	3	(2/5) 0,40	(3/10) 0,30
O que é o fogo de São Telmo?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	4	(4/5) 0,80	(7/10) 0,70
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	5	(4/5) 0,80	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	4	(4/5) 0,80	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	não	4	1	(2/5) 0,40	(2/10) 0,20
O que são os iaques?	sim	1	5	(4/5) 0,80	(6/10) 0,60
média	90,91%	1,18	3,86	67,27%	48,13%

Tabela A.30: Definição: frequência de termos + proximidade aos mais activos em espectro combinado

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	4	(3/5) 0,60	(4/10) 0,40
O que é a Brabançonne?	sim	1	5	(4/5) 0,80	(4/10) 0,40
O que é uma cítara?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que era o A6M Zero?	sim	1	4	(4/5) 0,80	(7/10) 0,70
O que é a paella?	sim	1	4	(3/5) 0,60	(3/10) 0,30
O que é a açorda?	sim	1	3	(2/5) 0,40	(7/10) 0,70
O que é o feta?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é o jagertee?	sim	1	2	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	5	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(2/5) 0,40	(3/10) 0,30
O que era a RSFSR?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é um berimbau?	sim	1	4	(3/5) 0,60	(7/10) 0,70
O que é VRML?	sim	1	4	(4/5) 0,80	(7/10) 0,70
O que são os forcados?	não	2	1	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	9	(5/5) 1,00	(8/10) 0,80
O que é a Feplam?	sim	1	3	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	6	(5/5) 1,00	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	6	(5/5) 1,00	(5/10) 0,50
O que é o Gil Vicente FC?	sim	1	2	(2/5) 0,40	(2/10) 0,20
O que é a Torre do Tombo?	sim	1	2	(3/5) 0,60	(3/10) 0,30
O que são os iaques?	sim	1	6	(5/5) 1,00	(7/10) 0,70
média	95,45%	1,05	3,95	65,45%	48,59%

Tabela A.31: Definição: frequência de termos + pesquisa na *Web*

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
O que é um menir?	sim	1	2	(4/5) 0,80	(4/10) 0,40
O que é a Brabançonne?	sim	1	3	(3/5) 0,60	(4/10) 0,40
O que é uma cítara?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que era o A6M Zero?	sim	1	7	(5/5) 1,00	(7/10) 0,70
O que é a paella?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é a açorda?	sim	1	4	(3/5) 0,60	(6/10) 0,60
O que é o feta?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é o jagertee?	não	2	1	(1/5) 0,20	(1/6) 0,17
O que é o ICCROM?	sim	1	3	(4/5) 0,80	(6/10) 0,60
O que é o IPM em Portugal?	sim	1	2	(1/5) 0,20	(3/10) 0,30
O que era a RSFSR?	sim	1	2	(2/5) 0,40	(5/10) 0,50
O que é um berimbau?	sim	1	5	(4/5) 0,80	(7/10) 0,70
O que é VRML?	sim	1	6	(5/5) 1,00	(8/10) 0,80
O que são os forcados?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é o fogo de São Telmo?	sim	1	3	(2/5) 0,40	(2/10) 0,20
O que é que é um brigadeiro?	sim	1	4	(4/5) 0,80	(8/10) 0,80
O que é a Feplam?	sim	1	2	(2/5) 0,40	(2/9) 0,22
O que é um kilt?	sim	1	5	(4/5) 0,80	(8/10) 0,80
O que é o Crescente Fértil?	sim	1	2	(3/5) 0,60	(5/10) 0,50
O que é o Gil Vicente FC?	não	3	1	(1/5) 0,20	(2/10) 0,20
O que é a Torre do Tombo?	sim	1	2	(3/5) 0,60	(4/10) 0,40
O que são os iaques?	sim	1	8	(5/5) 1,00	(7/10) 0,70
média	90,91%	1,14	3,73	64,55%	51,31%

Tabela A.32: Definição: frequência de termos + presença de nomes

Categoria Localização

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	39	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	15	1	(0/5) 0,00	(0/10) 0,00
Em que país fica a Ossétia do Norte?	não	26	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	18	1	(0/5) 0,00	(0/10) 0,00
Onde fica Livorno?	não	6	1	(0/5) 0,00	(2/10) 0,20
Onde é que José Eduardo Pinto da Costa nasceu?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	29	1	(0/5) 0,00	(0/10) 0,00
Onde é que Joana de Arc foi queimada?	não	10	1	(0/5) 0,00	(1/10) 0,10
Faro fica em que estado brasileiro?	não	30	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	3	1	(2/5) 0,40	(4/10) 0,40
Onde fica a Disneyland?	não	7	1	(0/5) 0,00	(2/10) 0,20
Em que país fica o Muro de Berlim?	não	33	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	5	(4/5) 0,80	(5/10) 0,50
Onde fica Porto Santo?	não	19	1	(0/5) 0,00	(0/10) 0,00
Em que concelho fica São Bento do Cortiço?	não	6	1	(0/5) 0,00	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	sim	1	2	(1/5) 0,20	(1/10) 0,10
média	27,78%	13,67	1,44	12,22%	12,22%

Tabela A.33: Localização: semelhança de espectro com bonificação crescente

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	33	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	14	1	(0/5) 0,00	(0/10) 0,00
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Onde é que José Eduardo Pinto da Costa nasceu?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	14	1	(0/5) 0,00	(0/10) 0,00
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	20	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	3	1	(2/5) 0,40	(6/10) 0,60
Onde fica a Disneyland?	não	2	1	(2/5) 0,40	(3/10) 0,30
Em que país fica o Muro de Berlim?	não	26	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	6	(5/5) 1,00	(6/10) 0,60
Onde fica Porto Santo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	38,89%	8,33	1,72	24,44%	18,89%

Tabela A.34: Localização: semelhança de espectro com bonificação limitada a 30

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	36	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	2	1	(1/5) 0,20	(1/10) 0,10
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	17	1	(0/5) 0,00	(0/10) 0,00
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(3/10) 0,30
Onde é que José Eduardo Pinto da Costa nasceu?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	22	1	(0/5) 0,00	(0/10) 0,00
Onde é que Joana de Arc foi queimada?	não	5	1	(1/5) 0,20	(1/10) 0,10
Faro fica em que estado brasileiro?	não	24	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	3	1	(2/5) 0,40	(6/10) 0,60
Onde fica a Disneyland?	não	2	1	(2/5) 0,40	(3/10) 0,30
Em que país fica o Muro de Berlim?	não	31	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	6	(5/5) 1,00	(5/10) 0,50
Onde fica Porto Santo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	33,33%	9,83	1,67	23,33%	16,67%

Tabela A.35: Localização: semelhança de espectro com bonificação limitada a 45

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	30	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	não	2	1	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	não	4	1	(1/5) 0,20	(1/10) 0,10
Onde vivem as aves tucanos?	não	16	1	(0/5) 0,00	(0/10) 0,00
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	16	1	(0/5) 0,00	(0/10) 0,00
Onde fica Livorno?	não	6	1	(0/5) 0,00	(1/10) 0,10
Onde é que José Eduardo Pinto da Costa nasceu?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	17	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	6	1	(0/5) 0,00	(3/10) 0,30
Onde fica a Disneyland?	não	2	1	(1/5) 0,20	(1/10) 0,10
Em que país fica o Muro de Berlim?	não	17	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	5	(4/5) 0,80	(5/10) 0,50
Onde fica Porto Santo?	não	6	1	(0/5) 0,00	(2/10) 0,20
Em que concelho fica São Bento do Cortiço?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	5	1	(1/5) 0,20	(1/10) 0,10
média	27,78%	8,72	1,44	13,33%	12,22%

Tabela A.36: Localização: proximidade aos nós mais activos do espectro combinado

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	23	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	12	1	(0/5) 0,00	(0/10) 0,00
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	13	1	(0/5) 0,00	(0/10) 0,00
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Onde é que José Eduardo Pinto da Costa nasceu?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	2	1	(1/5) 0,20	(1/10) 0,10
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	13	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	4	1	(2/5) 0,40	(3/10) 0,30
Onde fica a Disneyland?	não	2	1	(1/5) 0,20	(4/10) 0,40
Em que país fica o Muro de Berlim?	não	11	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	6	(5/5) 1,00	(6/10) 0,60
Onde fica Porto Santo?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	38,89%	6,44	1,72	22,22%	16,67%

Tabela A.37: Localização: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 15

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	30	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	não	2	1	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	14	1	(0/5) 0,00	(0/10) 0,00
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	15	1	(0/5) 0,00	(0/10) 0,00
Onde fica Livorno?	não	3	1	(1/5) 0,20	(1/10) 0,10
Onde é que José Eduardo Pinto da Costa nasceu?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	16	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	4	1	(1/5) 0,20	(3/10) 0,30
Onde fica a Disneyland?	não	2	1	(1/5) 0,20	(1/10) 0,10
Em que país fica o Muro de Berlim?	não	15	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	5	(4/5) 0,80	(6/10) 0,60
Onde fica Porto Santo?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Em que concelho fica São Bento do Cortiço?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	3	1	(1/5) 0,20	(1/10) 0,10
média	38,89%	7,56	1,61	20,00%	13,33%

Tabela A.38: Localização: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 30

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	30	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	não	2	1	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Onde vivem as aves tucanos?	não	16	1	(0/5) 0,00	(0/10) 0,00
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	15	1	(0/5) 0,00	(0/10) 0,00
Onde fica Livorno?	não	5	1	(1/5) 0,20	(1/10) 0,10
Onde é que José Eduardo Pinto da Costa nasceu?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	17	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	6	1	(0/5) 0,00	(3/10) 0,30
Onde fica a Disneyland?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que país fica o Muro de Berlim?	não	17	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	5	(4/5) 0,80	(5/10) 0,50
Onde fica Porto Santo?	não	4	1	(2/5) 0,40	(2/10) 0,20
Em que concelho fica São Bento do Cortiço?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	4	1	(1/5) 0,20	(1/10) 0,10
média	38,89%	8,22	1,56	16,67%	12,78%

Tabela A.39: Localização: proximidade aos nós mais activos do espectro combinado, limite de bonificação: 45

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	31	1	(0/5) 0,00	(0/10) 0,00
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	11	1	(0/5) 0,00	(0/10) 0,00
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	13	1	(0/5) 0,00	(0/10) 0,00
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Onde é que José Eduardo Pinto da Costa nasceu?	não	3	1	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	2	1	(1/5) 0,20	(1/10) 0,10
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	19	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	4	1	(2/5) 0,40	(3/10) 0,30
Onde fica a Disneyland?	não	2	1	(1/5) 0,20	(4/10) 0,40
Em que país fica o Muro de Berlim?	não	25	1	(0/5) 0,00	(0/10) 0,00
Onde fica São Bento do Ameixial?	sim	1	5	(4/5) 0,80	(6/10) 0,60
Onde fica Porto Santo?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	33,33%	8,06	1,50	21,11%	16,67%

Tabela A.40: Localização: semelhança de espectro + espectro combinado

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	9	1	(0/5) 0,00	(1/10) 0,10
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Onde é que José Eduardo Pinto da Costa nasceu?	não	3	1	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	6	1	(0/5) 0,00	(1/10) 0,10
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	17	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	4	1	(1/5) 0,20	(5/10) 0,50
Onde fica a Disneyland?	sim	1	3	(2/5) 0,40	(4/10) 0,40
Em que país fica o Muro de Berlim?	não	5	1	(1/5) 0,20	(1/10) 0,10
Onde fica São Bento do Ameixial?	sim	1	7	(5/5) 1,00	(6/10) 0,60
Onde fica Porto Santo?	sim	1	3	(2/5) 0,40	(3/10) 0,30
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	44,44%	4,61	1,83	25,56%	21,11%

Tabela A.41: Localização: confirmação por pesquisa *Web* + semelhança de espectro

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Em que distrito fica Chaves?	não	9	1	(0/5) 0,00	(1/10) 0,10
De onde é nativo o cisne branco?	sim	1	2	(1/5) 0,20	(1/5) 0,20
Onde fica o parque Eduardo VII?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Onde vivem as aves tucanos?	não	12	1	(0/5) 0,00	(0/10) 0,00
Em que país fica a Ossétia do Norte?	não	25	1	(0/5) 0,00	(0/10) 0,00
Em que ilha fica Sapporo?	não	4	1	(1/5) 0,20	(1/10) 0,10
Onde fica Livorno?	sim	1	2	(1/5) 0,20	(4/10) 0,40
Onde é que José Eduardo Pinto da Costa nasceu?	não	3	1	(2/5) 0,40	(2/10) 0,20
Em que país se realizaram os Jogos Olímpicos de 1976?	não	2	1	(1/5) 0,20	(1/10) 0,10
Onde é que Joana de Arc foi queimada?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Faro fica em que estado brasileiro?	não	15	1	(0/5) 0,00	(0/10) 0,00
Onde é a Igrejinha?	não	4	1	(2/5) 0,40	(3/10) 0,30
Onde fica a Disneyland?	não	3	1	(3/5) 0,60	(4/10) 0,40
Em que país fica o Muro de Berlim?	não	4	1	(1/5) 0,20	(1/10) 0,10
Onde fica São Bento do Ameixial?	sim	1	5	(4/5) 0,80	(6/10) 0,60
Onde fica Porto Santo?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Em que concelho fica São Bento do Cortiço?	não	2	1	(1/5) 0,20	(1/10) 0,10
Azaruja fica em que distrito de Portugal?	não	2	1	(1/5) 0,20	(1/10) 0,10
média	33,33%	5,06	1,56	25,56%	18,33%

Tabela A.42: Localização: confirmação por pesquisa *Web* + proximidade a nós mais activos do espectro combinado

Categoria Quantidade

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quantos jogadores tem um time de basquete?	não	104	1	(0/5) 0,00	(0/10) 0,00
Qual o diâmetro de Ceres?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Quantos focos tem uma elipse?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Que profundidade atinge a Fossa das Marianas?	sim	1	7	(5/5) 1,00	(7/10) 0,70
Qual a população da Ilha de Moçambique?	não	13	1	(0/5) 0,00	(0/10) 0,00
Quanto custou a Mars Observer?	sim	1	3	(2/3) 0,67	(2/3) 0,67
Qual a altura do Kebnekaise?	sim	1	6	(5/5) 1,00	(5/8) 0,62
Quantos ossos têm a face?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Com que idade o Meguinho foi campeão brasileiro de xadrez?	não	2	1	(2/5) 0,40	(2/6) 0,33
Quantos habitantes tinha Berlim em 1850?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Qual é a temperatura do zero absoluto?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Quantas faixas tem a bandeira dos Estados Unidos?	não	2	1	(1/5) 0,20	(1/10) 0,10
Quantos filmes realizou Jean Vigo?	não	3	1	(1/5) 0,20	(1/10) 0,10
Qual o comprimento da Ponte do Øresund?	sim	1	3	(2/5) 0,40	(4/10) 0,40
Quantas esposas tinha Ngungunhane?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Quantos metros saltou Nelson Évora nos Jogos Olímpicos?	sim	1	2	(2/5) 0,40	(2/8) 0,25
média	68,75%	8,44	2,38	39,17%	29,84%

Tabela A.43: Quantidade: afinidade semântica 1

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quantos jogadores tem um time de basquete?	não	104	1	(0/5) 0,00	(0/10) 0,00
Qual o diâmetro de Ceres?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Quantos focos tem uma elipse?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Que profundidade atinge a Fossa das Marianas?	sim	1	6	(5/5) 1,00	(7/10) 0,70
Qual a população da Ilha de Moçambique?	não	13	1	(0/5) 0,00	(0/10) 0,00
Quanto custou a Mars Observer?	sim	1	3	(2/3) 0,67	(2/3) 0,67
Qual a altura do Kebnekaise?	sim	1	4	(3/5) 0,60	(5/8) 0,62
Quantos ossos têm a face?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Com que idade o Mequinho foi campeão brasileiro de xadrez?	não	2	1	(2/5) 0,40	(2/6) 0,33
Quantos habitantes tinha Berlim em 1850?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Qual é a temperatura do zero absoluto?	sim	1	3	(3/5) 0,60	(3/5) 0,60
Quantas faixas tem a bandeira dos Estados Unidos?	não	2	1	(1/5) 0,20	(1/10) 0,10
Quantos filmes realizou Jean Vigo?	não	2	1	(1/5) 0,20	(1/10) 0,10
Qual o comprimento da Ponte do Øresund?	sim	1	3	(3/5) 0,60	(4/10) 0,40
Quantas esposas tinha Ngungunhane?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Quantos metros saltou Nelson Évora nos Jogos Olímpicos?	sim	1	2	(2/5) 0,40	(2/8) 0,25
média	68,75%	8,38	2,25	37,92%	31,09%

Tabela A.44: Quantidade: pesquisa na *Web*

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quantos jogadores tem um time de basquete?	não	108	1	(0/5) 0,00	(0/10) 0,00
Qual o diâmetro de Ceres?	sim	1	2	(3/5) 0,60	(3/5) 0,60
Quantos focos tem uma elipse?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Que profundidade atinge a Fossa das Marianas?	sim	1	9	(5/5) 1,00	(8/10) 0,80
Qual a população da Ilha de Moçambique?	não	13	1	(0/5) 0,00	(0/10) 0,00
Quanto custou a Mars Observer?	sim	1	3	(2/3) 0,67	(2/3) 0,67
Qual a altura do Kebnekaise?	sim	1	6	(5/5) 1,00	(5/8) 0,62
Quantos ossos têm a face?	sim	1	2	(1/5) 0,20	(2/10) 0,20
Com que idade o Mequinho foi campeão brasileiro de xadrez?	sim	1	2	(2/5) 0,40	(2/6) 0,33
Quantos habitantes tinha Berlim em 1850?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Qual é a temperatura do zero absoluto?	sim	1	3	(3/5) 0,60	(3/5) 0,60
Quantas faixas tem a bandeira dos Estados Unidos?	não	2	1	(1/5) 0,20	(1/10) 0,10
Quantos filmes realizou Jean Vigo?	não	2	1	(1/5) 0,20	(1/10) 0,10
Qual o comprimento da Ponte do Øresund?	sim	1	3	(3/5) 0,60	(4/10) 0,40
Quantas esposas tinha Ngungunhane?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Quantos metros saltou Nelson Évora nos Jogos Olímpicos?	sim	1	2	(2/5) 0,40	(2/8) 0,25
média	75,00%	8,56	2,62	40,42%	31,72%

Tabela A.45: Quantidade: afinidade + pesquisa na *Web*

Categoria Quando

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quando foi fundado o Nacional da Madeira?	não	2	1	(1/5) 0,20	(1/10) 0,10
Quando foi registado pela primeira vez o cometa Halley?	não	20	1	(0/5) 0,00	(0/10) 0,00
Quando se realizaram os Jogos Olímpicos de Munique?	não	5	1	(1/5) 0,20	(1/10) 0,10
Em que dia ocorreu a batalha de La Lys?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Quando é que foi assinada a Lei Áurea?	sim	1	2	(2/5) 0,40	(2/5) 0,40
Em que ano foi fundada a Bertrand?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Quando nasceu Machado de Assis?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Quando foi inaugurado o metropolitano de Lisboa?	não	3	1	(1/5) 0,20	(1/10) 0,10
Quando começa o outono no hemisfério sul?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que período existiu a República de Veneza?	não	8	1	(0/5) 0,00	(1/10) 0,10
Em que ano houve um terramoto no Irão?	não	18	1	(0/5) 0,00	(0/10) 0,00
Quando começou o Neolítico?	não	2	1	(2/5) 0,40	(2/10) 0,20
Quando nasceu Thomas Mann?	sim	1	2	(2/5) 0,40	(3/10) 0,30
Quando reinou Isabel II de Castela?	não	34	1	(0/5) 0,00	(0/10) 0,00
Em que ano foi construída a sinagoga de Curaçao?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Quando foi assinado o Tratado de Zamora?	não	3	1	(2/5) 0,40	(3/10) 0,30
Quando é que viveu Zenão de Eleia?	não	2	1	(1/5) 0,20	(1/8) 0,12
Quando foi fundado o Vasco da Gama?	não	3	1	(1/5) 0,20	(1/10) 0,10
Quando nasceu Vasco da Gama?	sim	1	2	(2/5) 0,40	(2/10) 0,20
média	36,84%	5,74	1,42	25,26%	15,39%

Tabela A.46: Quando: afinidade semântica 1

PERGUNTA	CORRECTA	IPRC	IPRE	ACERTO Top5	ACERTO Top10
Quando foi fundado o Nacional da Madeira?	não	2	1	(1/5) 0,20	(1/10) 0,10
Quando foi registado pela primeira vez o cometa Halley?	não	18	1	(0/5) 0,00	(0/10) 0,00
Quando se realizaram os Jogos Olímpicos de Munique?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Em que dia ocorreu a batalha de La Lys?	não	2	1	(1/5) 0,20	(2/10) 0,20
Quando é que foi assinada a Lei Áurea?	sim	1	3	(2/5) 0,40	(2/5) 0,40
Em que ano foi fundada a Bertrand?	sim	1	3	(2/5) 0,40	(2/10) 0,20
Quando nasceu Machado de Assis?	sim	1	2	(2/5) 0,40	(2/10) 0,20
Quando foi inaugurado o metropolitano de Lisboa?	não	6	1	(0/5) 0,00	(1/10) 0,10
Quando começa o outono no hemisfério sul?	não	2	1	(2/5) 0,40	(2/10) 0,20
Em que período existiu a República de Veneza?	não	9	1	(0/5) 0,00	(1/10) 0,10
Em que ano houve um terramoto no Irão?	não	18	1	(0/5) 0,00	(0/10) 0,00
Quando começou o Neolítico?	não	2	1	(2/5) 0,40	(2/10) 0,20
Quando nasceu Thomas Mann?	sim	1	2	(1/5) 0,20	(3/10) 0,30
Quando reinou Isabel II de Castela?	não	34	1	(0/5) 0,00	(0/10) 0,00
Em que ano foi construída a sinagoga de Curação?	sim	1	2	(1/5) 0,20	(1/10) 0,10
Quando foi assinado o Tratado de Zamora?	sim	1	3	(2/5) 0,40	(3/10) 0,30
Quando é que viveu Zenão de Eleia?	não	2	1	(1/5) 0,20	(1/8) 0,12
Quando foi fundado o Vasco da Gama?	não	4	1	(1/5) 0,20	(1/10) 0,10
Quando nasceu Vasco da Gama?	sim	1	3	(2/5) 0,40	(3/10) 0,30
média	42,11%	5,63	1,63	22,11%	15,92%

Tabela A.47: Quando: confirmação por pesquisa na *Web*

Bibliografia

[ACF+08] Carlos Amaral, Adán Cassan, Helena Figueira, André F. T. Martins, Afonso Mendes, Pedro Mendes, José Pina, and Cláudia Pinto. Priberam's question answering system in qa@clef 2008. In *CLEF Workshop*, pages 337–344, 2008.

[ACQS05] Isa Alves, Rove Chishman, Paulo Quaresma, and José Saias. Busca e extração de informações através de pergunta e resposta: uma nova concepção de web. In Unisinos, editor, *XXV Congresso da Sociedade Brasileira de Computação - III Workshop em Tecnologia da Informação e da Linguagem Humana*, pages 2218–2228, Brasil, July 2005.

[ALG01] Eugene Agichtein, Steve Lawrence, and Luis Gravano. Learning search engine specific query transformations for question answering. In *In Proceedings of WWW10*, pages 169–178, 2001.

[And83] John R. Anderson. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22:261–295, 1983.

[AP95] J.J. Almeida and Ulisses Pinto. Jspell – um módulo para análise léxica genérica de linguagem natural. In *Actas do X Encontro da Associação Portuguesa de Linguística*, pages 1–15, 1995.

[BDB02] Eric Brill, Susan Dumais, and Michele Banko. An analysis of the askmsr question-answering system. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 257–264. Association for Computational Linguistics, 2002.

[Bic00] Eckhard Bick. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

[Bil04] Matthew W. Bilotti. Query expansion techniques for question answering. Master's thesis, Massachusetts Institute of Technology, 2004.

[BRSS08] António Branco, Lino Rodrigues, João Silva, and Sara Silveira. Xisquê: An online qa service for portuguese. In *Computational Processing of the*

Portuguese Language, volume 5190 of *LNAI*, pages 232–235. Springer, 2008.

[Bus45] Vannevar Bush. As we may think. *The Atlantic Monthly*, July 1945.

[CA05] Silviu Cucerzan and Eugene Agichtein. Factoid question answering over unstructured and structured web content. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

[CC08] Luís Fernando Costa and Luís Miguel Cabral. Answering portuguese questions. In *Computational Processing of the Portuguese Language*, volume 5190 of *LNAI*, pages 228–231. Springer, 2008.

[CCL01] Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. Exploiting redundancy in question answering. In *In Proceedings of SIGIR*, pages 358–365. ACM Press, 2001.

[CdMR08] Gracinda Carvalho, David Martins de Matos, and Vitor Rocio. Idsay: Question answering for portuguese. Technical report, QA@CLEF 2008 - Working Notes, 2008.

[Cha09] Marcirio Silveira Chaves. *Uma Metodologia para Construção de Geo-Ontologias*. PhD thesis, Faculty of Sciences, University of Lisbon, 2009.

[CMG+08] Luísa Coheur, Ana Mendes, João Guimarães, Nuno J. Mamede, and Ricardo Ribeiro. Qa@l2f, second steps at qa@clef. Technical report, QA@CLEF 2008 - Working Notes, 2008.

[Cos08] Luís Costa. Esfinge at clef 2008: Experimenting with answer retrieval patterns. can they help? Technical report, QA@CLEF 2008 - Working Notes, 2008.

[CQ69] Allan Collins and M. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8, 1969.

[Cre97] Fabio Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11:453–482, 1997.

[DMP07] W. Duch, P. Matykiewicz, and J. Pestian. *Towards Understanding of Natural Language: Neurocognitive Inspirations*, volume 4669, chapter II, page 953–962. Springer Lecture Notes in Computer Science, 2007.

[DW05] Tiphaine Dalmas and Bonnie Webber. Using information fusion for open domain question answering. In *Proceedings of KRAQ 2005 Workshop, IJCAI*, 2005.

[DW07] Tiphaine Dalmas and Bonnie L. Webber. Answer comparison in automated question answering. *Journal of Applied Logic*, 5:104–120, 2007. Issue 1: Questions and Answers: Theoretical and Applied Perspectives.

[Esc08] Sergio Ferrández Escámez. *Arquitectura multilingüe de sistemas de búsqueda de respuestas basada en ILI y Wikipedia*. PhD thesis, Universidad de Alicante, Abril 2008.

[FEIBFEVG08] Óscar Ferrández Escamez, Rubén Izquierdo Beviá, Sergio Ferrández Escamez, and José Luis Vicedo González. Un sistema de búsqueda de respuestas basado en ontologías, implicación textual y entornos reales. *Procesamiento del lenguaje Natural*, 41, 2008.

[FPA+08] Pamela Forner, Anselmo Peñas, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe, and Erik Sang. Overview of the clef 2008 multilingual question answering track. In Francesca Borri, Alessandro Nardi, and Carol Peters, editors, *Cross Language Evaluation Forum: Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[GBG59] L. Green, E. Berkeley, and C. Gotlieb. Conversation with a computer. *Computers and Automation*, 1959.

[GFH+08] Danilo Giampiccolo, Pamela Forner, Jesús Herrera, Anselmo Peñas, Christelle Ayache, Dan Cristea, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu, and Richard Sutcliffe. Overview of the clef 2007 multilingual question answering track. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum - CLEF 2007, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*, pages 200–236, Budapest, Hungary, 09/2008 2008. Springer.

[GS04] Mark A. Greenwood and Horacio Saggion. A pattern based approach to answering factoid, list and definition questions. In *In Proceedings of the 7th RIAO Conference (RIAO 2004*, pages 26–28, 2004.

[GWCL61] Jr. Bert F. Green, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: an automatic question-answerer. In *IRE-AIEE-ACM '61 (Western): Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224. ACM, 1961.

[HGH+00] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin yew Lin. Question answering in webclopedia. In *In Proceedings of the Ninth Text REtrieval Conference (TREC-9*, pages 655–664, 2000.

[HHR02] E.H. Hovy, U. Hermjakob, and D. Ravichandran. A question/answer typology with surface text patterns. In *Proceedings of the DARPA Human Language Technology conference (HLT)*, 2002.

[Hu06] Haiqing Hu. *A Study on Question Answering System Using Integrated Retrieval Method*. PhD thesis, The University of Tokushima, Japan, 2006.

[Ing92] Peter Ingwersen. *Information Retrieval Interaction*. Taylor Graham Publishing, 1992. ISBN: 0-947568-549.

[Kai08] Michael Kaisser. The qualim question answering demo: supplementing answers with paragraphs drawn from wikipedia. In *HLT '08: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*. Association for Computational Linguistics, 2008.

[Kat88] Boris Katz. Using english for indexing and retrieving. In *Proceedings of the 1st RIAO Conference on User-Oriented Content-Based Text and Image Handling (RIAO '88)*, 1988.

[Kat97] Boris Katz. Annotating the world wide web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*, 1997.

[KEW01] Cody Kwok, Oren Etzioni, and Daniel S. Weld. Scaling question answering to the web. *ACM Transactions on Information Systems*, 19(3):242–262, 2001.

[KF93] Hideki Kozima and Teiji Furugori. Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, Netherlands, 1993. ISBN:90-5434-014-2.

[KFY+02] Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, Jimmy Lin, Gregory Marton, Alton Jerome McFarland, and Baris Temelkuran. Omnibase: Uniform access to heterogeneous data for question answering. In *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems* -, 2002.

[KL88] Boris Katz and Beth Levin. Exploiting lexical regularities in designing natural language systems. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-1988)*, Budapeste, Hungria, 1988.

[KLL+03] Boris Katz, Jimmy Lin, Daniel Loreto, Wesley Hildebr, Matthew Bilotti, Sue Felshin, Aaron Fern, Gregory Marton, and Federico Mora. Integrating web-based and corpus-based techniques for question answering. In *Proceedings of the Twelfth Text REtrieval Conference (TREC)*, pages 426–435, 2003.

[KR93] Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer, Dordrecht, 1993.

[KSN07] Jeongwoo Ko, Luo Si, and Eric Nyberg. A probabilistic framework for answer selection in question answering. In *In Proceedings of NAACL/HLT*, 2007.

[Kup93] Julian Kupiec. Murax: a robust linguistic approach for question answering using an on-line encyclopedia. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 181–190. ACM, 1993.

[Lam99] Sydney M. Lamb. *Pathways of the Brain: The Neurocognitive Basis of Language*. Amsterdam Studies in the Theory and History of Linguistic Science. John Benjamins North America, 1999. ISBN-55619-886-8.

[Leh77] Wendy Lehnert. *The process of Question Answering*. PhD thesis, Yale University, 1977.

[Lin02] Chin-Yew Lin. The effectiveness of dictionary and web-based answer reranking. In *Proceedings of the 19th international conference on Computational linguistics*. Association for Computational Linguistics, 2002.

[LK03] Jimmy Lin and Boris Katz. Question answering techniques for the world wide web. In *In Tutorial presentation at the European Chapter of the Association of Computational Linguistics - EACL*, 2003.

[LS04] H. Liu and P. Singh. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4), October 2004. Kluwer Academic Publishers.

[Lur76] A. R. Luria. *Cognitive Development: its Cultural and Social Foundations*. Harvard University Press, 1976. ISBN: 0-674-13732-9.

[LVF02] Fernando Llopis, José Luis Vicedo, and Antonio Ferrández. Passage selection to improve question answering. In *COLING-02: Proceedings of the 2002 Conference on Multilingual Summarization and Question Answering*. Association for Computational Linguistics, 2002.

[May03] Mark T. Maybury. Toward a question answering roadmap. In *New Directions in Question Answering*, pages 8–11, 2003.

[McC98] John McCarthy. What is artificial intelligence?, January 1998.

[Mil95] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 11:39–41, 1995.

[MNPT02a] Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. Is it the right answer?: exploiting web redundancy for answer validation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 425–432. Association for Computational Linguistics, 2002.

[MNPT02b] Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. Mining the web to validate answers to natural language questions. In *Proc. of the 3rd Int'l Conf. on Data Mining*, 2002.

[Mon03] Christof Monz. *From Document Retrieval to Question Answering*. PhD thesis, Institute for Logic, Language and Computation - University of Amsterdam, 2003. ISBN 90-5776-116-5.

[MRS09] C. D. Manning, P. Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2009.

[Niu07] Yun Niu. *Analysis of Semantic Classes: Toward Non-Factoid Question Answering*. PhD thesis, University of Toronto, 2007.

[OSGS08] Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes, and Nuno Seco. Papel: A dictionary-based lexical ontology for portuguese. In *Computational Processing of the Portuguese Language*, volume 5190 of *LNAI*, pages 31–40. Springer, 2008. ISBN: 978-3-540-85979-6.

[Per88] Luís Moniz Pereira. Inteligência artificial - mito e ciência. *Colóquio/Ciências - Revista de Cultura Científica*, 3:1–13, Outubro 1988. Fundação Calouste Gulbenkian.

[PRSV08] Anselmo Peñas, Alvaro Rodrigo, Valentín Sama, and Felisa Verdejo. Testing the reasoning for question answering validation. In *Journal of Logic and Computation*, volume 18 of *Special Issue on Natural Language and Knowledge Representation*, pages 459–474, June 2008.

[PVSV06] Rodrigo Peñas, Alvaro Verdejo, Valentín Sama, and Felisa Verdejo. Overview of the answer validation exercise 2006. In *Working Notes of the CLEF 2006 Workshop*, Alicante, Spain, Setembro 2006.

[PVV07] Rodrigo Peñas, Alvaro Verdejo, and Felisa Verdejo. Overview of the answer validation exercise 2007. In *Working Notes of the CLEF 2007 Workshop*, Budapeste, Hungria, Setembro 2007.

[QQR+04] Paulo Quaresma, Luis Quintano, Irene Rodrigues, José Saias, and Pedro Salgueiro. The university of evora approach to qa@clef-2004. In Carol Peters, editor, *Question-Answering Track of the Cross Language Evaluation Forum 2004*, Bath, UK, September 2004.

[QR05] Paulo Quaresma and Irene Rodrigues. A logic programming-based approach to qa@clef-2005. Technical report, Universidade de Évora, Question-Answering Track of the Cross Language Evaluation Forum - Vienna, Austria, 2005.

[RB05] Juliano Rabelo and Flávia Barros. Pergunte!: Uma interface em português para pergunta-resposta na web. In *Anais do XXV Congresso da Sociedade Brasileira de Computação*, volume 1, pages 1114–1117, São Leopoldo, 2005. UNISINOS.

[RJ97] C S. E. Robertson and K. Spärck Jones. Simple proven approaches to text retrieval. Technical Report 356, University of Cambridge Computer Laboratory, 1997.

[Rod07] Lino M. S. Rodrigues. Infra-estrutura de um serviço online de resposta-a-perguntas com base na web portuguesa. Master's thesis, Faculdade de Ciências da Universidade de Lisboa, 2007.

[RSGR75] Christopher K. Riesbeck, Roger C. Schank, Neil M. Goldman, and Charles J. Rieger, III. Inference and paraphrase by computer. *Journal of the ACM*, 22(3):309–328, 1975.

[RWJ+95] S. Robertson, S. Walker, S. Jones, M. Hancock-beaulieu, and M. Gatford. Okapi at trec-3. In *Proceedings of TREC-3*, pages 109–126, 1995.

[SAD08] Mark D. Smucker, James Allan, and Blagovest Dachev. Human question answering performance using an interactive information retrieval system. Technical report, Center for Intelligent Information Retrieval, DoCS, University of Massachusetts Amherst, Janeiro 2008.

[Sai03] José Saias. Uma metodologia para a construção automática de ontologias e a sua aplicação em sistemas de recuperação de informação. Mestrado em engenharia informática, Universidade de Évora, Novembro 2003.

[Sar03] Matthew Sargaison. Named entity information extraction and semantic indexing - techniques for improving automated question answering systems. Master's thesis, University of Sheffield, 2003.

[Sar06] Luis Sarmento. Siemes - a named-entity recognizer for portuguese relying on similarity rules. In *PROPOR 2006 - Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, Rio de Janeiro - Brasil, 2006.

[Sim65] R. F. Simmons. Answering english questions by computer: a survey. *Communications of the ACM*, 8(1):53–70, 1965.

[Sou01] M. Soubbotin. Patterns of potential answer expressions as clues to the right answers. In *In Proceedings of the Tenth Text REtrieval Conference - TREC*, pages 293–302, 2001.

[SPC06] Luís Sarmento, Ana Sofia Pinto, and Luís Cabral. Repentino - a wide-scope gazetteer for entity recognition in portuguese. In *Computational Processing of the Portuguese Language: 7th International Workshop (PROPOR 2006), Brasil*, volume LNAI 3960, pages 31–40. Springer Verlag, 2006.

[SQ05] José Saias and Paulo Quaresma. A methodology to create legal ontologies in a logic programming information retrieval system. In V.R. Benjamins et al., editor, *Law and the Semantic Web*, volume 3369 of *LNAI*, pages 185–200. Springer-Verlag, 2005. ISBN: 3-540-25063-8.

[SQ06] José Saias and Paulo Quaresma. A proposal for an ontology supported news reader and question-answer system. In Solange Oliveira Rezende and Antonio Carlos Roque da Silva Filho, editors, *2nd Workshop on Ontologies and their Applications (WONTO'06) in the Proceedings of International Joint Conference, 10th Ibero-American Artificial Intelligence Conference, IBERAMIA-SBIA-SBRN*, Ribeirão Preto, Brazil, October 2006. ICMC-USP. ISBN: 85-87837-11-7.

[SQ07] José Saias and Paulo Quaresma. The senso question answering approach to portuguese qa@clef-2007. Technical report, Universidade de Évora, Cross-Language Evaluation Forum Workshop, Budapest, Hungary, 2007. ISSN: 1818-8044, ISBN: 2-912335-32-9.

[SQ08a] José Saias and Paulo Quaresma. The senso question answering system at qa@clef 2008 - working notes. Technical report, Universidade de Évora, 2008.

[SQ08b] José Saias and Paulo Quaresma. The university of Évora's participation in qa@clef-2007. In C. Peters, V. Jijkoun, Th. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras, and D. Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval - 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science, Sublibrary: Information Systems and Applications, incl. Internet/Web, and HCI*, pages 316–323. Springer-Verlag, 2008. ISBN: 978-3-540-85759-4.

[STO08] Luís Sarmento, Jorge Teixeira, and Eugénio Oliveira. Experiments with query expansion in the raposa (fox) question answering system. Technical report, QA@CLEF 2008 - Working Notes, 2008.

[Tur50] A. M. Turing. Computing machinery and intelligence. *Mind*, 1950.

[TVA07] George Tsatsaronis, Michalis Vazirgiannis, and Ion Androutsopoulos. Word sense disambiguation with spreading activation networks generated from thesauri. In *20th International Joint Conference in Artificial Intelligence - IJCAI 2007*, Hyderabad, India, Janeiro 2007.

[Van95] Lucretia Vanderwende. *The Analysis of Noun Sequences using Semantic Information Extracted from On-Line Dictionaries*. PhD thesis, Georgetown University, October 1995.

[VI90] Jean Veronis and Nancy Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th conference on Computational linguistics*, Helsinki, Finland, 1990.

[VPV08] Alvaro Verdejo, Rodrigo Peñas, and Felisa Verdejo. Overview of the answer validation exercise 2008. In *Working Notes of the CLEF 2008 Workshop*, Aarhus, Denmark, Setembro 2008.

[Wei66] Joseph Weizenbaum. Eliza: A computer program for the study of natural language communication between man and machine. *Communciations of the ACM*, 9(1), 1966.

[Woo73] W. A. Woods. Progress in natural language understanding: an application to lunar geology. In *AFIPS '73: Proceedings of the June 4-8, 1973, national computer conference and exposition*, pages 441–450. ACM, 1973.

[Woo77] W.A. Woods. Lunar rocks in natural english: Explorations in natural language question answering. In A. Zampolli, editor, *Linguistic Structures Processing*, pages 521–569. North-Holland, 1977.

[Wu03] Lynn Wu. Prism: An answer projection system. Master's thesis, Massachusetts Institute of Technology, 2003.

[Yus08] Alvaro Rodrigo Yuste. Validación de respuestas: Metodología de evaluación y desarrollo de sistemas. Master's thesis, Universidad Nacional de Educación a Distancia, 2008.

[Zhe03] Zhiping Zheng. Answerbus news engine. In *The Twelfth International World Wide Web Conference (WWW 2003)*, Hungary, May 2003.

Índice