ESFINGE - RESPOSTA A PERGUNTAS USANDO A REDE

Luís Fernando Costa
Linguateca at SINTEF ICT
Pb 124 Blindern, 0314 Oslo, Noruega
luis.costa@sintef.no

RESUMO

Este artigo começa por dar uma breve panorâmica da área da resposta automática a perguntas. Referem-se alguns dos sistemas disponíveis na Rede e alguns dos eventos que procuram medir os avanços na área. De seguida descreve-se o sistema de resposta automática a perguntas em português Esfinge e os resultados obtidos até ao momento pelo mesmo. Terminam-se com algumas considerações sobre a utilidade de sistemas de resposta automática a perguntas e o estado da arte nesta área para o português.

PALAVRAS-CHAVE

Resposta automática a perguntas, Processamento de linguagem natural, Recolha de informação, Rede (Web), N-gramas, Padrões.

1. INTRODUÇÃO

Os motores de busca mais populares (como o Google, o Yahoo ou o MSN) procuram documentos relacionados com um conjunto de palavras-chave fornecidas pelos utilizadores. Ao invés, um sistema de resposta automática a perguntas tenta responder com precisão a questões formuladas em linguagem natural.

Alguns exemplos de sistemas de resposta automática a perguntas acessíveis na Rede são o AnswerBus (http://www.answerbus.com) e o Brainboost (http://www.brainboost.com). Estes sistemas respondem a questões de domínio geral, mas outros sistemas há que respondem a questões num domínio específico. O ExtrAns (http://www.ifi.unizh.ch/CL/extrans/) é um exemplo: este sistema responde a questões sobre UNIX.

Desde 1999 que se realizam eventos para tentar medir os avanços nesta área. Por exemplo as tarefas de Question Answering do TREC (http://trec.nist.gov/data/qa.html) principalmente para o inglês e do CLEF (http://clef-qa.itc.it/) que privilegia a multilingualidade e outras línguas para além do inglês (o português nomeadamente passou a ser uma das línguas do CLEF a partir de 2004) [Santos & Rocha, 2005].

Tal como descrito em [Silva, 2003] para os motores de busca e em [Santos, 1999] para o processamento de linguagem natural, muito trabalho específico há a fazer para cada linguagem em particular. Para o português existe um sistema de resposta a perguntas acessível na Rede criado pela Priberam (http://www.priberam.pt/trust/publico.aspx). A Universidade de Évora tem também um sistema de resposta a perguntas em português que participou no CLEF tanto na edição de 2004 como na de 2005 [Quaresma et al, 2005]. Para além destes existe o sistema Esfinge que descreverei brevemente neste artigo.

2. DESCRIÇÃO DO SISTEMA

Segundo as mitologias egípcia e grega, a Esfinge era um demónio que vivia perto de Tebas e que apresentava enigmas às pessoas que a encontravam. Quem não fosse capaz de decifrar os enigmas era estrangulado sem contemplações. O nome do sistema de resposta a perguntas Esfinge é inspirado neste violento mito da Antiguidade.

O Esfinge baseia-se na arquitectura proposta por Eric Brill [Brill, 2003], que defende a possibilidade de obter bons resultados, aplicando técnicas simples a grandes quantidades de informação. O sistema está descrito detalhadamente em [Costa, 2005a, 2005b].

Ao receber uma pergunta em linguagem natural, o Esfinge começa por convertê-la em padrões de respostas plausíveis. Por exemplo, para a pergunta *Onde fica Paredes de Coura?*, um dos padrões gerados seria *"Paredes de Coura" fica* (as aspas indicam que as palavras dentro das mesmas devem aparecer seguidas).

De seguida pesquisam-se estes padrões na Rede (até ao momento invocando o motor de pesquisa Google para executar essa tarefa). Obtêm-se dessa forma um conjunto de excertos dos documentos seleccionados pelo Google onde se espera encontrar respostas à pergunta. Visto que se verificou em experiências anteriores, que alguns tipos de sítios podem comprometer a qualidade das respostas obtidas, criou-se uma lista de padrões de endereços a não serem considerados, lista esta que inclui padrões tais como *blog*, *humor* e *piadas*. Esta lista foi criada manualmente, mas em experiências futuras poder-se-ão usar técnicas mais sofisticadas para classificar as páginas Web [Aires et al, 2005] .

O passo seguinte consiste em extrair N-gramas de palavras (de comprimento 1 a 3) - ou seja sequências de 1 a 3 palavras - dos primeiros 100 excertos de documentos obtidos no módulo anterior. Estes N-gramas serão posteriormente ordenados de acordo com a seguinte fórmula:

Pontuação de um N-grama = $\sum$ (F * P * C), nos 100 primeiros excertos obtidos da pesquisa na Rede; em que F é a frequência do N-grama, P a pontuação do padrão utilizado para recuperar o documento de onde o N-grama foi extraído e C o comprimento do N-grama.

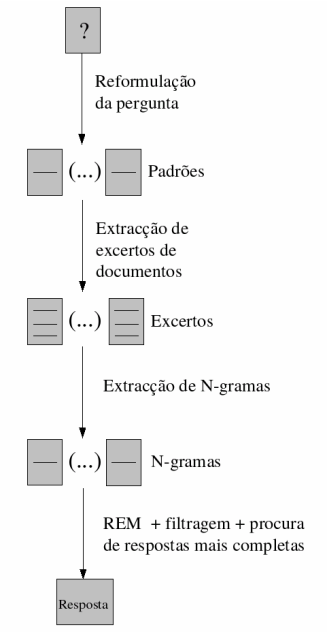

Figura 1. Arquitectura do Esfinge

Por vezes, o tipo de pergunta pode ser fulcral para a procura da resposta. Por exemplo, uma pergunta começando pela palavra *Quando* sugere que a resposta muito provavelmente será uma data. O Esfinge tem um módulo que usa o reconhecedor/classificador de entidades mencionadas SIEMES para tentar detectar entidades mencionadas dos tipos desejados. O SIEMES [Sarmento et al, 2005] detecta entidades mencionadas num vasto leque de categorias, das quais o Esfinge usa um subconjunto para os seus propósitos. As categorias consideradas são *Pessoa*, *País*, *Localidade* (inclui nomes de cidades, vilas, etc.), *Localização geográfica* (localizações sem conotação política directa, como por exemplo *África*), *Data* e *Quantidade*.

Nos casos em que o tipo de pergunta sugere um ou mais tipos de resposta, enviam-se ao SIEMES os 200 N-gramas de palavras com pontuação mais elevada obtidos dos módulos anteriores. Caso se encontrem entidades mencionadas dos tipos pretendidos, a pontuação atribuída a essas respostas é incrementada.

Seguidamente a lista das respostas possíveis é passada por um conjunto de filtros:

- Um filtro que rejeita palavras contidas nas questões. Por exemplo a palavra *Amália* não é desejada como resposta para a pergunta *Quem foi Amália?*
- Um filtro que rejeita respostas contidas numa lista de "respostas indesejáveis". As palavras nesta lista são palavras muito frequentes que não respondem a perguntas isoladamente (alguns exemplos são *pessoas*, *nova*, *lugar* e *grandes*).
- Um filtro que usa o analisador morfológico jspell [Simões & Almeida, 2002] para determinar as categorias gramaticais das palavras que constituem as potenciais respostas. Este filtro exclui as respostas cujas primeira e última palavras não são nomes comuns ou próprios, adjectivos ou números. Desta forma conseguem-se excluir palavras não interessantes neste contexto, como preposições ou interjeições.

A resposta final será a resposta candidata com melhor pontuação que conseguir passar todos os filtros referidos anteriormente. No entanto, se existir outra resposta que inclua a resposta com melhor pontuação e não seja rejeitada por nenhum filtro, essa resposta mais comprida será a resposta retornada pelo sistema.

O Esfinge pode ser acedido através do URL http://www.linguateca.pt/Esfinge/ e está instalado num Pentium 4 - 2,4 GHz, com 1 GB de memória RAM e o sistema operativo Red Hat Linux 9. Na sua versão actual, o sistema demora de 1 a 2 minutos a responder às questões que lhe são apresentadas.

3. RESULTADOS

O Esfinge participou na tarefa de resposta a perguntas do CLEF em 2004 e 2005 (http://www.linguateca.pt/CLEF/). Nesta tarefa os sistemas participantes recebem um conjunto de 200 perguntas preparadas pela organização e uma colecção de documentos. Os sistemas devem retornar a resposta a cada uma das perguntas acompanhada do código de um documento que a justifique. As perguntas são maioritariamente do tipo factóide (ex: *Quem é o presidente do Brasil?*), mas existem também definições (ex: *Quem é Brad Pitt?*) e perguntas restritas temporalmente (*Quem era o guarda-redes titular do Benfica em 1979?*) foram adicionadas em 2005.

Na tabela 1 apresentam-se os resultados obtidos pelo Esfinge nos CLEF 2004 e 2005. Graças a essa participação foram detectadas e corrigidas algumas deficiências no sistema. Na tabela apresentam-se também os resultados obtidos pela versão actual do sistema com as perguntas de 2004 e 2005 e os resultados do melhor sistema (U. Amsterdam) e do melhor sistema para o português (Universidade de Évora) em 2004 e 2005 (em que o sistema da Priberam obteve os melhores resultados entre todos os sistemas).

Tabela 1. Resultados do CLEF 2004 e 2005.

	Sistema	Nº perguntas	# Respostas certas	% Respostas certas
CLEF 2004	Esfinge	199	30	15%
	Melhor sistema para o português	199	56	28%
	Melhor sistema	200	91	46%
	Esfinge actual	199	55	28%
CLEF 2005	Esfinge	200	48	24%
	Melhor sistema	200	129	65%
	Esfinge actual	200	61	31%

Resposta(s) do Esfinge

Mon Sep 5 14:31:15 CEST 2005

Pergunta: ***Mencione um escritor norueguês.***

Jostein Gaarder

Ganhe um livro autografado por Jostein Gaarder! (Noruega - o site Ganhe um livro autografado por Jostein Gaarder! O renomado escritor norueguês Jostein Gaarder acaba de lançar no Brasil o livro A Garota das Laranjas.

Mais TV O escritor norueguês Jostein Gaarder (autor do best-seller O Mundo de Sofia) e os Segundo a crítica, Érico Veríssimo foi um escritor que soube,

Jostein Gaarder - Wikipédia Jostein Gaarder é um escritor norueguês nascido no dia 8 de agosto de 1952. Ele é autor de romances, contos e histórias infantis. Gaarder nasceu em Oslo,

Gaarder

Ganhe um livro autografado por Jostein Gaarder! (Noruega - o site Ganhe um livro autografado por Jostein Gaarder! O renomado escritor norueguês Jostein Gaarder acaba de lançar no Brasil o livro A Garota das Laranjas.

Mais TV O escritor norueguês Jostein Gaarder (autor do best-seller O Mundo de Sofia) e os Segundo a crítica, Érico Veríssimo foi um escritor que soube,

Jostein Gaarder - Wikipédia Jostein Gaarder é um escritor norueguês nascido no dia 8 de agosto de 1952. Ele é autor de romances, contos e histórias infantis. Gaarder nasceu em Oslo,

Figura 2. Respostas do Esfinge

O Esfinge pode ser experimentado livremente pela Rede. A figura 2 mostra algumas das respostas sugeridas pelo sistema para a questão *Mencione um escritor norueguês*. Além das respostas, o sistema

fornece também ligações para alguns dos documentos de onde as respostas foram extraídas. Desta forma, o utilizador pode confirmar se a resposta foi ou não satisfatória, ou encontrar informação mais detalhada relacionada com a questão formulada.

4. CONCLUSÃO

Por vezes os utilizadores procurando informação pretendem obter respostas exactas e não listas de documentos fastidiosas. Nestes casos um sistema de resposta a perguntas pode dar uma resposta mais satisfatória do que um motor de busca. Não existem ainda muitos sistemas desenvolvidos para o português. Com o desenvolvimento e a participação do sistema Esfinge no CLEF pretende-se dar um contributo ao avanço desta área ainda pouco explorada.

O trabalho planeado para o futuro inclui explorar os resultados que se conseguem obter investindo no pré-processamento de corpora, testar padrões mais complexos (considerando posições e distâncias) e implementar uma base de dados para armazenar/consultar perguntas já respondidas.

Um dos objectivos da apresentação do Esfinge na presente conferência é aumentar o seu leque de utilizadores, permitindo a compilação de perguntas reais que possibilitem não só testar melhor o sistema mas também torná-lo mais útil para os seus utilizadores.

AGRADECIMENTO

Gostava de agradecer a Diana Santos pela revisão de versões anteriores deste artigo. Este trabalho é financiado pela Fundação para a Ciência e Tecnologia através do projecto POSI/PLP/43931/2001, co-financiada pelo POSI.

REFERÊNCIAS

Aires, R., Aluísio S. and Santos, D, 2005. User-aware page classification in a search engine. *Proceedings of Stylistic Analysis Of Text For Information Access, SIGIR 2005 Workshop.* Salvador, Brasil.

Brill, E., 2003. Processing Natural Language without Natural Language Processing. *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics.*Mexico City, pp. 360-9.

Costa, L., 2005. First Evaluation of Esfinge - a Question Answering System for Portuguese. *Advances in Cross-Language Information Retrieval: Fifth Workshop of the Cross-Language Evaluation Forum.*Bath, Reino Unido.

Costa, L., 2005. 20th Century Esfinge (Sphinx) solving the riddles at CLEF 2005. *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop.*Viena, Áustria. (no prelo)

Quaresma P., Quintano, L., Rodrigues, I., Saias J., and Salgueiro P., 2005. The University of Évora approach to QA@CLEF-2004. *Advances in Cross-Language Information Retrieval: Fifth Workshop of the Cross-Language Evaluation Forum.*Bath, Reino Unido.

Santos, D., 1999. Towards language-specific applications. *In Machine Translation*, Vol. 2, No. 14, pp 83-112.

Santos D. and Rocha, P., 2005. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. *Advances in Cross-Language Information Retrieval: 5th Workshop of the Cross-Language Evaluation Forum.*Bath, UK.

Sarmento, L., Pinto, A. S. and Cabral, L., 2005. REPENTINO – A collaborative wide-scope gazetteer for Entity Recognition in Portuguese. (enviado para apreciação)

Silva, Mário J., 2003. The Case for a Portuguese Web Search Engine. *Actas IADIS International Conference WWW Internet do XVII Encontro da Associação Portuguesa de Linguística.* Algarve, Portugal.

Simões, A. M. and Almeida, J.J, 2002. Jspell.pm - um módulo de análise morfológica para uso em Processamento de Linguagem Natural. *Actas do XVII Encontro da Associação Portuguesa de Linguística.* Lisboa, pp. 485-495.