Do tópico às respostas: do processo humano à sua simulação

Luísa Coheur
INESC-ID/IST
luisa.coheur@inesc-id.pt

Ângela Costa
INESC-ID/UNL
angela@l2f.inesc-id.pt

Resumo

No quadro do projecto de uma disciplina de Língua Natural, 8 grupos de alunos participaram no Págico tendo como objectivos: a) identificar os processos envolvidos na procura das respostas aos tópicos; b) identificar técnicas, recursos linguísticos ou ferramentas que poderiam ser úteis na automatização desses processos. Este artigo resume e discute as metodologias apresentadas e os elementos que poderiam ser usados para as implementar, numa tentativa de compreender o que pode, efectivamente, ser realizado por uma máquina.

Palavras chave

Pesquisa de tópicos, Técnicas de Processamento de Língua Natural, Recursos linguísticos

1 Introdução

Não é de todo trivial identificar com exactidão as etapas realizadas por um humano na sua pesquisa de respostas a um dado tópico[1]. No entanto, a identificação destas etapas pode ser de extrema utilidade, pois estas representam potenciais passos a implementar numa máquina com os mesmos objectivos. Assim sendo, e tendo em conta a tarefa a realizar no Págico, foi proposto a um conjunto de 23 pessoas que participasse nesta competição, mas mais do que procurar exaustivamente as páginas relevantes foi-lhes pedido que tentassem **abstrair** as diferentes estratégias levados a cabo com o objectivo de as encontrar. Mais ainda, foi-lhe posto como meta que identificassem, dentro dos **recursos disponíveis** para as comunidades ligadas ao Processamento de Língua Natural (PLN), os que poderiam ser usados por uma máquina com o objectivo de automatizar essas estratégias. Essas 23 pessoas frequentavam à data da competição a disciplina de Língua Natural do Mestrado em Engenharia Informática e de Computadores, do Instituto Superior Técnico (Tagus Park), e este foi um dos projectos em que trabalharam no quadro dessa cadeira. Cada grupo (num total de 8 grupos) ficou de responder a um conjunto específico de tópicos, sendo o cardinal desse conjunto definido em função do número de elementos do grupo (em média cada aluno ficou responsável por sete questões). O que se descreve neste artigo representa uma reflexão tendo por base os relatórios entregues[2], apresentando-se e discutindo-se as estratégias referidas de modo explícito ou implícito pelos diferentes grupos. Apesar de serem muito variados os pontos destacados nos diferentes trabalhos, a metodologia geral de procura é comum e engloba duas "grandes" etapas: formulação da *query* (sequência de termos a submeter ao motor de pesquisa) e análise de resultados.

A organização deste artigo é a seguinte: na secção 2 discute-se a metodologia geral de pesquisa, na secção 3 discute-se a etapa que leva à formulação da *query* a submeter ao motor de pesquisa e na secção 4 o foco vai para a análise dos documentos devolvidos e escolha das respostas. Na secção 5 apresentam-se referências a trabalho relacionado e, finalmente, na secção 6, são tiradas as principais conclusões, apontando-se ainda para trabalho futuro.

2 Metodologia geral

2.1 Da *query* para os textos e destes para novas *queries*

Em traços largos, vários grupos referem uma abordagem de "tentativa e erro", isto é, formulam uma *query* – obtida, de algum modo, a partir do tópico (ver secção 3) – analisam os resultados e, caso não encontrem uma resposta nos tex-

[1] A palavra *tópico* tem aqui o mesmo significado atribuído no Págico, isto é, a sequência de palavras que representa a informação a pesquisar.

[2] Os relatórios encontram-se em http://www.inesc-id.pt/ficheiros/publicacoes/8124.pdf.

tos devolvidos pelo motor de pesquisa, nem em *links* que ocorram nestes textos (ver secção 4), reformulam a *query* e voltam a repetir o processo. A escolha dos termos a usar na *query* é de extrema importância, dado que deles depende a qualidade dos resultados devolvidos pelo motor de pesquisa. Estes termos podem ser usados numa única *query* ou lançados em *queries* paralelas, sendo a informação das páginas encontradas cruzada na procura de resposta. Por exemplo, as respostas ao tópico [jornais portugueses que existiam no tempo da implantação da república] foram obtidas com base em informação presente nas páginas devolvidas perante as *queries* `jornais portugueses` e `implantação da república`. De modo semelhante, alguns termos são usados para encontrar páginas e outros para navegar nestas e chegar à resposta. Por exemplo, dado o tópico [Escritores Lusófonos traduzidos para 5 ou mais idiomas], a *query* é formulada com os termos `Escritores Lusófonos` e o termo `traduz` usado para encontrar a resposta na(s) página(s) devolvida(s).

De notar que em todo este processo há um *know-how*, difícil de quantificar, que já se tinha ou que se vai ganhando sobre um dado tópico e que permite realizar *queries* cada vez mais sofisticadas; uma implementação capaz de simular este ganho, obrigaria a combinar extracção de informação com técnicas de raciocínio.

2.2 Quando considerar esgotada uma fonte de informação?

O facto de se saber que podem existir tópicos sem resposta, leva a que mesmo em termos humanos seja difícil decidir quando parar uma pesquisa. Ou seja, não é possível garantir que uma pesquisa não alcançou resultados porque a *query* estava mal formulada ou, simplesmente, porque a informação não existe. Automatizar este processo é algo que está longe de ser resolvido e é um tema muito interessante de investigação. Um grupo sugere o factor tempo como o decisor, num processo que teria igualmente em conta a complexidade da *query*. No entanto, avaliar a complexidade de uma *query* não é fácil e decidir um limite para o tempo no qual se tem de encontrar uma resposta é algo extremamente subjectivo.

2.3 Consulta de fontes de informação externas

O problema anterior leva-nos à questão do Págico ser uma fonte de informação limitada, no sentido em que é um subconjunto da Wikipédia Portuguesa. Alguns grupos, considerando que, de algum modo, a pesquisa através do motor do Págico não traria resultados, recorreram-se de fontes de informação externa quer para encontrar respostas aos tópicos, quer para refinar a escolha dos termos na sua pesquisa. Na verdade, dado que o objectivo do projecto, tal como referido, estava mais focado na abstracção do processo de pesquisa da resposta do que na obtenção de todas as respostas expectáveis no quadro do Págico, esta decisão era esperada. A Wikipédia Portuguesa, bem como a Inglesa foram assim fontes de informação alternativas, e o Google foi igualmente muito usado. Fontes de informação mais específicas foram também exploradas. Por exemplo, um grupo refere a consulta à revista da Forbes na pesquisa do tópico [Empresários Estrangeiros com fortuna considerável]. De notar que a escolha destas fontes de informação específica nunca foi identificada como uma etapa explícita em todo o processo. De facto, é algo que um humano faz naturalmente, tendo por base o conhecimento que tem do mundo, mas que não é fácil de automatizar.

3 Do tópico à formulação da *query*

Dado um tópico, a primeira etapa a realizar é a que culmina numa *query* a apresentar ao motor de pesquisa em causa. Este processo compreende várias fases que se ilustram de seguida.

3.1 Identificação dos termos do tópico

Na base da formulação das *queries* está o conhecimento que cada elemento tem do mundo. E, apesar de não ser referido por muitos grupos, não é óbvio que, por exemplo, dado o tópico [Eventos onde Maria de Lurdes Mutola foi medalha de ouro], a partição dos termos na formulação da *query* seja `Maria de Lurdes Mutola` e `medalha de ouro`. Ou seja, um humano, identifica claramente que `Maria de Lurdes Mutola` é o nome de uma pessoa e que `medalha de ouro` é um termo composto.

Apesar de este processo de partição ser trivialmente realizado por um humano e, daí, grande parte dos trabalhos não o referirem, um grupo em particular refere a utilização dos N-gramas mais frequentes para fazer a partição do tópico em termos, bem como de um reconhecedor de entidades mencionadas para a identificação de nomes de pessoas, locais, etc.; algum tipo de *chunker* seria igualmente de extrema utilidade para detecção de sequências como *medalha de ouro*.

3.2 Compreensão do tópico

Vários grupos descreveram explicitamente uma etapa de compreensão do tópico. No entanto, dado que normalmente a informação em causa num dado tópico é facilmente compreendida, só se torna notório que esta é uma tarefa a ter em conta em todo o processo quando a interpretação do tópico não é evidente.

3.2.1 Desconhecimento de termos

Em vários trabalhos é referido o desconhecimento de certas palavras de um tópico, o que levou os alunos a recorrerem-se de dicionários. Por exemplo, um grupo deparou-se com o tópico [Doenças que acometeram a população indígena nas América] e, tendo dúvidas sobre o significado exacto da palavra *acometer*, usaram dicionários para encontrar sinónimos. De notar que os sinónimos encontrados podem ser usados, numa fase posterior, na formulação da *query* a submeter (ver secção 3.4). Outro grupo explica que dado o tópico [Países que venceram a copa do mundo em uma disputa de penalties], tiveram de confirmar que o termo brasileiro *copa do mundo* se referia ao campeonato do mundo de futebol, pois o termo *penalties* não se refere exclusivamente à modalidade de futebol.

3.2.2 Termos/tópicos imprecisos

No entanto, mais do que apenas procurar significados de termos, vários grupos referem dificuldades na interpretação do tópico devido ao facto de estes conterem termos difíceis de quantificar. Por exemplo, a expressão *ligação forte* em [Pintores estrangeiros com uma ligação forte a Portugal ou ao Brasil] é difícil de definir. Neste contexto, o tópico [Jornais que circularam no Rio de Janeiro entre 1910 e 1960] pode levantar igualmente algumas dúvidas: não é claro se se refere a jornais apenas do Rio de Janeiro ou de outra parte qualquer que circularam apenas nesse período no Rio de Janeiro. O mesmo se passa com o tópico [Politicos lusófonos do século XX assassinados]. Deverão ser devolvidos os políticos que nasceram, foram assassinados ou viveram nesse século?

3.2.3 É realmente necessário compreender um tópico?

Embora não seja fácil interpretar alguns tópicos, também pode acontecer que as respostas encontradas venham resolver essa questão. Por exemplo, se em relação aos políticos lusófonos do século XX assassinados só surgirem pessoas nascidas no século XX e nesse século assassinados, a questão da interpretação deixa de ser problemática. No caso do tópico [Países que venceram a copa do mundo em uma disputa de pênalties], a formulação de uma *query* com as sequências `copa do mundo` e `penalties` iria resultar apenas em artigos com potenciais respostas. Ou seja, estas duas sequências, quando associadas, acabariam por eliminar resultados não relacionados. Este possibilidade de encontrar os resultados sem ter realmente necessidade de compreender o tópico em causa é de extrema importância na automatização do processo: a máquina não tem de compreender o significado dos termos para chegar às páginas com as respotas; tem apenas de ser capaz de escolher os termos certos para a sua pesquisa. Um exemplo que ilustra bem o facto de não ser necessário compreender o significado dos termos do tópico é o reportado por um grupo que se debateu com [Filmes sobre o cangaço]. Não tendo a mínima ideia do que significaria *cangaço*, começaram por submeter ao motor de pesquisa uma *query* usando os termos `filmes` e `cangaço`, não obtendo resultados; posteriormente, fizeram uma pesquisa apenas com o termo desconhecido, encontrando uma página que lhe era dedicada. Nesta página depararam-se com os nomes de vários *cangaceiros* famosos. Analisando as páginas destas personalidades, uma a uma, acabaram por chegar a filmes baseados nas suas vidas, encontrado assim respostas para o tópico em causa.

3.3 Identificar o tipo da *query*

Mais do que compreender o tópico de modo a formular a *query* correcta, há que saber o tipo de conhecimento que está envolvido para ser posteriormente capaz de escolher entre os resultados devolvidos os que satisfazem o tópico. Um dos grupos sugere um processo concreto de classificação, exemplificando com [Quem descobriu São Tomé e Principe?]; neste caso, o pronome interrogativo indica que a resposta terá de ser uma pessoa ou um grupo de pessoas. Se para um humano esta tarefa é praticamente óbvia, a sua automatização tem sido fruto de muita investigação (ver secção 5).

3.3.1 O "excesso" de informação

O outro lado da moeda, tal como referido por alguns grupos, diz respeito ao facto de algumas respostas serem previamente conhecidas pelos alunos. Nesses casos, as respostas eram usadas para formular a *query*, tal como ilustra o primeiro exemplo da secção 3.4.3. O que nos leva à secção

que se segue onde se discute como formular a *query* a submeter ao motor de pesquisa.

3.4 Formulação das *queries*

A formulação de *queries*, como seria de esperar, foi a etapa mais destacada em todos os trabalhos. Seguem-se as estratégias de formulação de *queries* identificadas pelos diferentes grupos, bem como de técnicas, recursos linguísticos e ferramentas que poderiam participar na implementação destas estratégias.

3.4.1 Eliminação de termos e de partes de termos

Apesar dos tópicos não serem exactamente perguntas completas em língua natural e serem normalmente de dimensões reduzidas, a prática de eliminação de termos foi seguida por todos os grupos. Assim, por exemplo, a *query* obtida a partir do tópico [Praias de Portugal boas para a prática de surf] seria `Praias surf Portugal`, ou seja, as palavras *de*, *boas*, *para*, *a* são removidas durante a formação da *query*. A implementação deste processo corresponderia à eliminação de palavras funcionais (e de alguns advérbios/adjectivos) e poderia ser implementada recorrendo a uma lista de *stopwords* e/ou etiquetadores morfo-sintácticos. Vários grupos referem a classificação morfo-sintática através de técnicas específicas – como por exemplo, usando HMMs – ou através da utilização de ferramentas como o Tree-Tagger (Schmid, 1994).

Outra prática amplamente sugerida diz respeito à eliminação de sufixos de palavras; os termos obtidos podem ser os lemas dos termos em causa ou apenas seus prefixos. Por exemplo, o grupo que ficou de responder ao tópico [Telenovelas brasileiras passadas no tempo da escravatura no Brasil], refere que usou nas suas pesquisas o prefixo `escrav`, tendo obtido pesquisas com as palavras *escravo*, *escrava*, *escravatura*, *escravidão*, etc.

São igualmente referidos casos em que é usado o singular em detrimento do plural (e mesmo do masculino em vez do feminino).

Com o objectivo de automatizar estes processos são referidos lematizadores e mesmo *stemmers* como o Porter Stemmer (Porter, 1980), sendo sugerida a sua extensão para Português.

3.4.2 Expansão básica de termos

A expansão de termos dos tópicos é talvez a mais destacada em todos os trabalhos. Nesta secção descrevem-se as técnicas sugeridas que seriam de (relativamente) fácil implementação, sendo a secção que se segue dedicada a expansões que já implicam raciocínios complexos e de difícil automatização.

A expansão de acrónimos é referida por um grupo que exemplifica estes casos com a palavra *FRELIMO* que é expandida para *Frente de Libertação de Monçambique*, termo usado na pesquisa. Existem *sites* onde se podem pesquisar acrónimos, incluindo a própria Wikipédia. De notar que os próprios documentos alcançados numa primeira pesquisa com o acrónimo podem trazer a informação necessária para que numa segunda pesquisa se possam usar os acrónimos expandidos. Este caso ilustra bem a interacção que existe entre os vários processos.

A utilização de relações semânticas como a sinonímia, hiperonímia e meronímia é igualmente amplamente referida. Um exemplo de utilização de sinónimos já foi referida anteriormente quando perante o tópico [Doenças que acometeram a população indígena nas América], os alunos foram procurar uma definição mais precisa da palavra *acometeram*; outro é o uso da palavra `povos` em vez de `tribos` perante o tópico [Tribos indígenas que vivem na Amazónia]. Quanto à utilização de hiperónimos um caso relatado consistiu na utilização das palavras `mamíferos` e `herbívoros` numa pesquisa que tinha no tópico a palavra *Zebra*. O uso de merónimos é também explícito na formação de *queries* com as expressões `políticos portugueses`, `políticos brasileiros`, `político moçambicanos`, etc., a partir de *políticos lusófonos*. Vários grupos referem a utilização de dicionários como o da Priberam[3] e da Wikipédia (Fellbaum, 1998), neste processo.

3.4.3 Expansão não trivial de termos

Um caso que ilustra bem como a formulação de *queries* feita por um humano pode ser difícil de reproduzir é o do tópico [Guitarristas portugueses que também foram compositores]. Nesta situação, os alunos lembraram-se logo do Carlos Paredes, pelo que a primeira *query* foi feita com o nome desse grande músico, ou seja, neste caso, conhecendo respostas possíveis ao tópico, o processo de pesquisa tratou-se apenas de encontrar páginas que validassem essas respostas. Esta estratégia, apesar de ter sido um caso isolado, mostra bem que existem recursos dos quais os humanos se podem recorrer (o seu conhecimento do mundo) e que são dificilmente implementáveis

[3] `http://www.priberam.pt/`.

(apenas a existência de uma base de dados de factos poderia simular esta abordagem).

Apesar do caso anterior ser um extremo, a utilização de termos resultantes de relações complexas entre palavras, bem como de raciocínios elaborados, são referidos em vários trabalhos. Neste contexto, é mencionada a utilização de paráfrases. Por exemplo, a formulação da *query* `estiveram presos` é criada como paráfrase da expressão *passaram temporadas na prisão* ocorrida no tópico; *toureiros a cavalo* origina `cavaleiros tauromáquico`. No entanto, outros exemplos relatados já não correspondem exactamente a paráfrases. Exemplos concretos – e ainda relativamente simples – são os pares *crianças*/`infantil` ou *ensino superior*/`faculdade`. Exemplos particularmente elaborados são os que apresentam o termo `biocombustível` obtido a partir de [Produtos agrícolas com os quais se pode produzir combustível em escala comercial], ou `história de Moçambique` a partir de [Personagens do século XX ligadas à luta anti-colonial em Moçambique]. Um outro caso interessante em que os alunos explicam o raciocínio que os levou a uma *query* bem sucedida, deu-se com o tópico [Mamíferos herbívoros existentes em Moçambique]. Depois de terem esgotado todas as hipóteses básicas de formulação de *queries* (`animais Moçambique`, `fauna Moçambique`,...) sem obter resultados, um dos elementos do grupo lembrou-se que uma amiga Moçambicana lhe costumava falar dos parques naturais que visitava. A pesquisa passou a ser feita com os termos `parques naturais` e `reservas naturais` e foram encontrados resultados.

Todos estes pontos ilustram bem como a expansão da *query* pode ter de ser feita com base em termos não habitualmente relacionados nos recursos disponíveis.

3.5 Escolha dos termos da *query* (e dos termos para navegação)

A escolha dos termos a submeter, sejam estes provenientes de modo directo do tópico ou resultado de algum tipo de expansão, básica ou complexa, é outra das tarefas não triviais, pois não é possível prever com exactidão se um dado conjunto de termos vai ser bem sucedido ou não (no caso desta competição, ainda mais difícil de prever era, dado a base de informação ser apenas um subconjunto de páginas da Wikipédia). Um exemplo interessante que ilustra bem este problema é o apresentado pelo grupo responsável pelo tópico [Cantores vaiados nos festivais de música brasileira na década de 60]. Dado que *query* `cantores vaiados` não obtinha resposta e que a *query* `década de 60` devolveria uma grande quantidade de resultados irrelevantes, a solução foi submeter a *query* `festival de música brasileira` e depois ir pesquisar os que tinham tido lugar na década de 60 (ou seja, nem todos os termos são usados na *query* submetida, sendo alguns "reservados" para a navegação nos resultados, tal como já referido e tal como explicado na secção 4). Ora um humano é capaz de compreender que algumas pesquisas (por exemplo, `década de 60`), não fazem sentido, pois são demasiado genéricas, mas é muito difícil implementar este processo de decisão numa máquina. Neste quadro, um dos grupos propõe uma estratégica mais definida, referindo as seguidas etapas que vão sendo percorridas se não se encontraram respostas (suficientes) na etapa anterior: a) a *query* é formulada com base em todos os termos do tópicos; b) são eliminadas preposições e os artigos; c) são eliminados adjectivos e verbos ou usam-se prefixos de termos.

Há aqui que referir (finalmente) uma vantagem da máquina nesta pesquisa: o formular e voltar a formular *queries* torna-se rapidamente uma tarefa penosa para um humano; uma máquina pode jogar com permutações de todos os termos que forem possíveis candidatos a (partes de) *queries*. Neste ponto, o limite de uma máquina pode estar bem mais à frente de um humano e tem apenas a ver com a sua capacidade de processamento.

3.6 *Queries* paralelas

Como já foi referido, algumas pesquisas são feitas em modo paralelo, sendo os resultados cruzados no fim. Ou seja, em vez de *queries* formuladas com todos os termos em vista, são escolhidos alguns para uma *query* e outros para outra (e eventualmente para mais), sendo os resultados cruzados no fim. Para além do exemplo já apresentado na secção 2, temos o caso da formulação das *queries* `documentários políticos` e `documentários brasileiros` de modo a encontrar a resposta a [documentários políticos brasileiros]. Mais uma vez, a escolha destes termos, é difícil de realizar por uma máquina.

4 Análise dos documentos e escolha dos resultados a apresentar

Após a inserção da *query* no motor de pesquisa é devolvido um conjunto de páginas (potenciais respostas do sistema), cabendo ao participante

escolher as que são realmente respostas ao tópico em questão. As técnicas usadas neste processo de análise são apresentadas de seguida.

4.1 Tópico como categoria da Wikipédia

O caso mais fácil de resolver, referido por todos os grupos, acontece quando o tópico ou a *query* formulada correspondem a categorias da Wikipédia (por exemplo, Frutos de Angola é uma categoria da Wikipédia, ou seja todos os frutos marcados com essa categoria serão resposta ao tópico [Frutos de Angola]). Se o tópico coincide exactamente com a categoria (raro), basta devolver todos os elementos dessa lista; caso contrário, há que verificar os elementos da lista de modo a escolher aqueles que verificam as restrições adicionais do tópico, não submetidas na *query*. Um dos grupos divide esta situação em dois casos: no primeiro a lista a percorrer é curta e fácil de percorrer (por exemplo, o que acontece com o tópico [Dinossauros carnívoros que habitaram o Brasil]); no segundo, em que a lista em causa é muito grande, torna-se complicado consultar todas as páginas devolvidas (por exemplo, o que sucede com o tópico [Filmes brasileiros sobre futebol] em que a pesquisa com os termos `Filmes Brasileiros` devolve uma extensa lista). Para este último caso, um dos grupos chegou a implementar um pequeno programa em *XQuery*[4] para facilitar essa pesquisa.

4.2 Caso geral

Infelizmente, nem sempre a pesquisa é assim tão fácil, ou seja, nem sempre termos dos tópicos coincidem com categorias da Wikipédia.

Nesta situação, os métodos de análise (ou navegação) nas páginas devolvidas multiplicam-se. Em traços gerais, dada uma página, são procurados nesta os tópicos ou termos usados nas pesquisas. Vários grupos referem o uso de técnicas que se assemelham às usadas na formulação de *queries* para navegar/localizar na(s) página(s) os pedaços de texto relevantes. Quando estes pedaços de informação são encontrados, o aluno detecta se contêm a resposta. Caso contenha, a página é devolvida; caso contrário, poderá encontrar-se na página um *link* a explorar, ou novos termos a usar numa futura pesquisa. De lembrar que vários grupos referem o cruzamento de informação de várias páginas.

Há que notar que todos estes processos, desde o decidir se a resposta se encontra num dado parágrafo ao optar por seguir um *link* (ou não), são tópico de investigação, por exemplo em sistemas de pergunta/resposta. De notar que a capacidade de descartar respostas erradas é feita (normalmente) sem dificuldade por um humano, mas não por uma máquina.

5 Trabalho Relacionado

A tarefa a realizar no Págico tem as suas raízes numa anterior competição denominada GikiClef[5], mais orientada para questões com restrições geográficas. As competições de sistemas de pergunta/resposta como as que têm lugar no quadro do CLEF[6] e do TREC[7] estão também relacionadas, apesar dos sistemas em competição lidarem usualmente com questões bem formadas em língua natural, e terem de devolver a resposta exacta às questões e não apenas a página. No entanto, os tópicos do Págico representam uma dificuldade acrescida, pois quase todos envolvem restrições complexas (`anos 60`, `medalha de ouro`, `traduzidos para 5 ou mais idiomas`, etc.), o que normalmente não acontece nas perguntas em jogo nas competições acima referidas. No que se segue, faz-se um breve paralelo entre os sistemas de pergunta/resposta e a tarefa a realizar no Págico.

5.1 Os sistemas de pergunta/resposta

Os sistemas de pergunta/resposta apresentam, tipicamente, três módulos: o primeiro responsável pela interpretação da pergunta e formulação da *query*; o segundo pela recuperação de informação onde se poderá encontrar as respostas; o terceiro pela selecção da resposta.

Na etapa dita de interpretação, é feita a classificação da pergunta com o objectivo de determinar o tipo esperado da resposta. São inúmeros os trabalhos que se dedicam à classificação da pergunta, não apenas fazendo variar técnicas (Li e Roth, 2002), (Huang, Thint e Qin, 2008), (Silva et al., 2011) como as taxonomias em causa (Hermjakob, Hovy e Lin, 2002), (Li e Roth, 2002). A formulação da *query*, incluindo as técnicas de expansão, é também alvo de muita investigação, (Brill, Dumais e Banko, 2002), (Wang et al., 2005), (Monz, 2011).

No que diz respeito à etapa de *retrieval*, é nesta que são encontrados os pedaços de texto, potenciais fontes de respostas. Estes textos provêm da Web ou de colecções de documentos. Existem também sistemas que pré-processam as

[4] `http://www.w3.org/TR/xquery/`.

[5] `http://www.linguateca.pt/GikiCLEF/`.

[6] `http://www.clef-initiative.eu//`.

[7] `http://trec.nist.gov/`.

fontes de informação, criando bases de conhecimento (Saias e Quaresma, 2007).

Finalmente, na etapa de selecção das respostas, são identificadas e escolhidas a(s) resposta(s) a devolver pelo sistema, sendo esta etapa igualmente alvo de muita investigação (ver (Mendes e Coheur, 2012) para um *survey* sobre *answering*).

5.2 Sistemas de pergunta/resposta vs. Págico

O primeiro e o terceiro módulos acima descritos (o módulo de interpretação e o de selecção da resposta) equivalem, em traços gerais, aos elementos descritos nas secções 3 e 4. O módulo de *retrieval* tem mais a ver com o motor de pesquisa em si o que, neste caso, estava limitado ao motor do Págico (apesar dos alunos terem referido, por exemplo, o uso do Google).

5.2.1 O fluxo de informação

Nos sistemas tradicionais de pergunta/resposta a informação obtida através de uma *query* não é usada para refinar uma *query* ou a própria navegação. Como se viu, este processo (muito difícil de implementar) é a base do trabalho realizado por humanos no quadro do Págico e é talvez o ponto mais complexo a simular nesta tarefa.

5.2.2 A etapa de interpretação e formulação da query

Apesar da classificação da questão ser uma tarefa fundamental nos sistemas de pergunta/resposta, pois como estes têm de devolver a resposta exacta à pergunta (e não um documento) a categoria da pergunta permite-lhes validar os candidatos a respostas – apenas os que correspondem à categoria esperada são devolvidos – só um grupo fala na importância de classificar *queries*. Tal poderá dever-se, exactamente, ao facto de grande parte dos trabalhos acima mencionados terem por alvo a classificação de questões e não de *queries*. Sendo as primeiras normalmente mais compridas e recorrendo-se usualmente de elementos como pronomes interrogativos que dão boas pistas, à priori, para o tipo da resposta, a classificação de *queries* é mais complexa. Adicionalmente, é também relativamente fácil para um humano decidir que um tópico como [Escritores Lusófonos traduzidos para 5 ou mais idiomas] tem como alvo pessoas pelo que este problema terá passado despercebido a grande parte dos alunos.

Quanto à formulação das *queries* as técnicas e recursos apresentados pelos diferentes grupos correspondem ao trabalho que se faz habitualmente nesta tarefa.

5.2.3 A selecção da resposta

No que diz respeito ao módulo de selecção de resposta dos sistemas de pergunta/resposta, este tem por missão identificar potenciais respostas e seleccionar uma ou mais entre várias candidatas. Neste contexto, a tarefa do Págico é, por um lado, mais simples, pois a página toda é devolvida, mas, por outro lado, mais complexa, pois tem de lidar com as restrições impostas no tópico.

6 Conclusões e Trabalho Futuro

Neste trabalho apresentou-se um apanhado das diferentes estratégicas realizadas por um conjunto de alunos que participaram no Págico, na sua tentativa de encontrar respostas aos tópicos pelos quais ficaram responsáveis; adicionalmente, foram sugeridos recursos que poderiam ajudar a implementar as referidas estratégias. Grande foco foi dado à formulação de *queries*. O modo como as respostas eram encontradas nas páginas também foi referido em todos os trabalhos. No entanto, algumas tarefas que os alunos levaram a cabo, apesar de fundamentais nestas pesquisas foram alvo apenas de destaque pontual, pois pelo facto de serem tão óbvias de realizar por um humano, poucos se aperceberam que faziam parte do processo de pesquisa.

De todo o trabalho apresentado, a capacidade de um humano em tirar informação de pesquisas mal conseguidas será talvez o ponto fulcral para o sucesso deste tipo de tarefas e é, sem dúvidas, o mais complicado de implementar. No processo de formulação de *queries* tornou-se também óbvio que uma pessoa é capaz de estabelecer relações semânticas invulgares entre palavras, o que lhe permite refinar as *queries* a submeter; uma máquina dificilmente estabeleceria "à primeira" essas relações. Interessante seria compreender como o Watson (Ferrucci et al., 2010) se comportaria neste tipo de competições, dado que é um dos poucos sistemas capazes de explorar ligações não triviais entre palavras.

Particularmente complicada de implementar é também a tarefa de identificar se um texto é ou não portador de uma resposta, em particular de modelar a capacidade de verificar se as restrições que fazem parte de praticamente todos os tópicos do Págico são satisfeitos ou não. Uma última nota para o facto de, podendo não existir resposta a um tópico, a decisão de quando parar

a pesquisa, não ser nada trivial.

Do lado da máquina, identifica-se apenas a vantagem de poder correr, sem esforço, inúmeras *queries*.

No geral, a tarefa proposta aos alunos resultou num projecto muito interessante, porque os obrigou a abstrair as pequenas tarefas executadas nas suas pesquisas, porque lhes permitiu participar numa avaliação conjunta e, finalmente, porque os obrigou a realizar uma ponte entre os processos que tinham em mãos e a matéria leccionada. Uma avaliação detalhada dos resultados obtidos está fora do âmbito deste trabalho, no entanto, a título de curiosidade, o grupo obteve o segundo lugar da participação humana. Dado que o foco estava realmente na metodologia para alcançar a resposta, muitas questões foram respondidas com páginas que não correspondiam realmente a uma resposta, mas em que esta se encontrava algures na página, sendo o campo das justificações usado para indicar porque é que tinha sido escolhida tal página. Desde modo perderam-se pontos importantes.

Quanto ao Págico, o facto de apresentar tópicos com restrições complexas faz com seja fácil compreender que um sistema, capaz de encontrar automaticamente as respostas em causa, facilitaria imensamente a pesquisa humana; ao contrário das perguntas normalmente presentes em competição de sistemas de pergunta/resposta, estes tópicos não se resolvem, como se viu, rapidamente, com um motor de pesquisa qualquer e na primeira tabela da Wikipédia que aparecer. Tarefas semelhantes serão certamente um dos grandes desafios para os próximos tempos.

Agradecimentos

Este trabalho teve o apoio da FCT, através de fundos do programa PIDDAC, e do projecto PT-STAR (CMU-PT/HuMach/0039/2008) que financia a bolsa da Ângela Costa. Agradecemos também aos alunos da disciplina de Língua Natural, MEIC-T, IST, cuja participação no Págico e constatações pertinentes nos seus relatórios serviram de base a este trabalho.

Referências

Brill, Eric, Susan Dumais, e Michele Banko. 2002. An analysis of the askmsr question-answering system. Em *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pp. 257–264, Morristown, NJ, USA. Association for Computational Linguistics.

Fellbaum, C., editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Ferrucci, David A., Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, e Christopher A. Welty. 2010. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79.

Hermjakob, Ulf, Eduard Hovy, e Chin-Yew Lin. 2002. Automated question answering in webclopedia: a demonstration. Em *Proceedings of the second international conference on Human Language Technology Research*, pp. 370–371, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Huang, Zhiheng, Marcus Thint, e Zengchang Qin. 2008. Question classification using head words and their hypernyms. Em *EMNLP*, pp. 927–936.

Li, Xin e Dan Roth. 2002. Learning question classifiers. Em *Proceedings of the 19th international conference on Computational linguistics*, pp. 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Mendes, Ana Cristina e Luísa Coheur. 2012. When the answer comes into question in question-answering: survey and open issues. *Natural Language Engineering*, January, 2012.

Monz, Christof. 2011. Machine learning for query formulation in question answering. *Natural Language Engineering*, 17(04):425–454.

Porter, M. F. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Saias, José e Paulo Quaresma. 2007. The university of évora's participation in qa@clef-2007. Em *CLEF*, pp. 316–323.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Em *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49.

Silva, João, Luísa Coheur, Ana Mendes, e Andreas Wichert. 2011. From symbolic to subsymbolic information in question classification. *Artificial Intelligence Review*, 35:137–154.

Wang, Yi-Chia, Jian-Cheng Wu, Tyne Liang, e Jason S. Chang. 2005. Web-based unsupervised learning for query formulation in question answering. Em *IJCNLP*, pp. 519–529.