O desafio da participação humana do IT-Coimbra no Págico

Arlindo Veiga
Instituto de Telecomunicações, Coimbra
DEEC - Universidade de Coimbra
aveiga@co.it.pt

Carla Lopes
Instituto de Telecomunicações, Coimbra
Instituto Politécnico de Leiria
calopes@co.it.pt

Dirce Celorico
Instituto de Telecomunicações, Coimbra
dircelorico@co.it.pt

Jorge Proença
Instituto de Telecomunicações, Coimbra
jproenca@co.it.pt

Fernando Perdigão
Instituto de Telecomunicações, Coimbra
DEEC - Universidade de Coimbra
fp@co.it.pt

Sara Candeias
Instituto de Telecomunicações, Coimbra
saracandeias@co.it.pt

Resumo

Na qualidade de grupo de investigação em Processamento Computacional da Língua portuguesa, pretendemos, neste documento, relatar a experiência vivenciada na participação do grupo *LudIT* no Págico – Português Mágico.

Estando o nosso trabalho mais centrado, de uma forma geral, no Processamento Automático da Fala, exprimimos obrigatoriamente uma visão decorrente de, como participantes humanos, ter entrado num desafio que levanta questões de língua distintas das que, até ao momento, têm sido levantadas no âmbito da investigação que temos desenvolvido e que estão mais relacionadas com o Processamento da Linguagem Natural. Num relato breve, descrevemos a estratégia adotada e as dificuldades encontradas. Decorrentes delas, apresentamos igualmente algumas opiniões, as quais podem vir a ser consideradas como sugestões a acolher numa próxima edição do Págico ou em outro desafio de perfil semelhante. Finalizamos com uma tentativa de interpretação do resultado obtido pela participação do *LudIT*.

Palavras chave

LudIT, Wikipédia, Participação Humana

1 O Porquê da Participação

No contexto da comunidade científica do processamento Computacional da Língua Portuguesa, será consensual admitir que o Processamento da Linguagem Natural, a Linguística Computacional e o Processamento da Fala são áreas que se encontram relacionadas e que a compreensão da estrutura da Língua Portuguesa com vista ao seu processamento passa também pelo entendimento quer das necessidades quer das dificuldades sentidas por cada uma dessas áreas. A possibilidade da participação humana no Págico foi encarada, no seio do grupo de investigação de Processamento da Fala do Instituto de Telecomunicações (polo de Coimbra), como uma primeira abordagem ao tema do processamento da linguagem natural e da recuperação de informação e como uma forma pormenorizada de entender a problemática da obtenção de respostas não triviais em arquivos de informação complexos. Acabou por se tornar um desafio cativante no sentido de conseguir responder, de forma tão completa quanto possível, às questões levantadas.

2 A Estratégia

A estratégia adotada na participação do *LudIT* no Págico começou pela divisão de trabalho pelos seus 6 elementos, tendo sido atribuído a cada elemento um conjunto equivalente de tópicos (uma média de 25 tópicos[1] por elemento). Inicialmente, foi utilizado o sistema SIGA (Costa, Mota e Santos, 2012), mas convergiu-se rapidamente para a pesquisa de temas através da Wikipédia on-line (Wikipédia, 2012). Assumimos que a grande maioria das páginas não teria sido atualizada desde abril de 2011 até à altura da nossa participação (novembro de 2011). Tal foi verificado na maioria dos casos, com apenas algumas exceções.

O processo de formulação de termos a subme-

[1] Adotamos a palavra *tópico* com o mesmo sentido atribuído pelo Págico, isto é, uma sequência de palavras que representa a informação a pesquisar.

ter ao motor de busca, de um modo generalizado, traduziu-se essencialmente pela identificação de palavras-chave. Como palavras-chave foram admitidas expressões compostas, tais como cristalizações (como no **tópico 039** [Jogos Olímpicos]) ou nomes próprios (no **tópico 095** [São Tomé e Príncipe]). Da mesma forma, foram por vezes utilizados, na pesquisa, lemas, isto é, *palavras* sem determinações de morfemas (ou desinências gramaticais, tal como o plural) ou de lexemas (ou desinências lexicais, tal como os sufixos): pesquisas por **tópico 146** [vulcão] em vez de [vulcões], por **tópico 104** [ordem religiosa] em vez de [ordens religiosas], ou por **tópico 136** [desporto] em vez de [desportivas], são disso exemplos. Em outras situações, as palavras-chave representaram expansões, como são exemplos a pesquisa por **tópico 010** [culinária do Brasil] em vez de [pratos brasileiros] ou por **tópico 153** [tauromaquia] em vez de [toureiros a cavalo]. A abstração necessária para chegar às palavras-chave implicou, naturalmente, uma interpretação da pergunta baseada num conhecimento complexo, com aportes linguísticos e culturais externos ao que está representado nos tópicos, mas quase imediato para um humano.

A procura na enciclopédia livre on-line mostrou-se eficiente (no sentido de retribuir alguns resultados e de, eles próprios, permitirem refinar a procura e a localização de outros) para dar respostas a ações de pergunta complexa, bem como se revelou muito rápida (frações de segundo) na devolução de resultados.

Em suma, a pesquisa por palavras-chave, bem como o algoritmo embebido no sistema de pesquisa da Wikipédia para dar respostas a partir de palavras parecidas, constituiu um fator decisivo nos resultados alcançados. A pesquisa por categoria, possível na Wikipédia on-line, também acelerou o processo de obtenção de páginas relevantes. Introduzindo, no SIGA, o título das páginas devolvidas pela Wikipédia on-line, foi sempre lá encontrada uma opção de resposta. Bastou então verificar se a informação da resposta existia nessa página da versão de abril de 2011.

3 As Dificuldades

Como participantes humanos, sentimos algumas contrariedades em ultrapassar certas dificuldades, principalmente as relacionadas com o elevado número de respostas a associar a um tópico. A título de exemplo, ultrapassava 50 o número de respostas corretas ligadas ao **tópico 019** [Tribos indígenas que vivem na Amazónia]. Seria talvez interessante devolver menos respostas, mas estar a elas associado um grau de importância ou de relevância no que ao tópico diz respeito. Por outro lado, foi também evidente a ausência de respostas na Wikipédia a algumas das questões. Ao **tópico 153** [Toureiros a cavalo de países lusófonos com carreira internacional], por exemplo, não pôde ficar associado nenhum dos cavaleiros tauromáquicos Ribeiro Telles, pelo facto de a sua atividade, ainda que claramente conhecida no meio tauromáquico, não vir suficientemente representada na Wikipédia on-line. Exemplos como este evidenciam a necessidade de que os conteúdos da Wikipédia, por forma a acautelarem uma representação de informação sociocultural e enciclopedista, devem ser continuadamente alargados.

No sistema SIGA, uma das dificuldades encontradas prendeu-se com o tempo de espera para obter as páginas quando o tema de pesquisa devolvia a uma lista muito extensa. Seria mais funcional apresentar um menor número de resultados, mas de maior relevância. O facto de a pesquisa por categorias, também devolvidas pelo SIGA, não se encontrar funcional, foi outra das causas que condicionou a utilização do sistema.

A ambiguidade gerada pela enunciação de algumas questões, apesar de ser esse o objetivo do desafio, foi outra das dificuldades sentidas no ato de selecionar respostas. Para indicar os **tópico 144** [Locais referidos n'Os Lusíadas], dever-se-iam considerar espaços geográficos como Continentes e Rios? E a ilha encantadora simbolizada pela Ilha dos Amores? E o cabo das tormentas figurado no Adamastor? E os **tópico 122** [Políticos lusófonos do século XX assassinados]? Teriam que ter nascido e, também, teriam que ter sido assassinados, na extensão do séc. XX? Ou seriam aceites respostas que assegurassem apenas uma das asserções?

Um outro aspeto muito revelador da dificuldade do desafio (nada trivial, de facto), é que a resposta estava, algumas vezes, dependente da interpretação textual, reclamando uma leitura interpretativa do conteúdo (vd. resposta ao **tópico 152** [Pintores estrangeiros com uma ligação forte a Portugal ou ao Brasil], como exemplo). Uma participação menos cuidada, ou uma máquina menos treinada, poderia levar a dar respostas sem sentido. Para terem sido consideradas como válidas as respostas *Joca (político)* e *Ênio Ricardo Gomes* ao **tópico 122** [Políticos lusófonos do século XX assassinados], a máquina deveria ter interpretado as relações sintáticas e semânticas existentes nas expressões complexas [quando ia para uma reunião com o então governador Marcello Alencar ele foi assassinado com

11 tiros] e [Ênio foi vitimado por um atentado a tiros], respetivamente.

O facto de, para responder a questões não triviais, ter exigido detetar focos (pontos ou palavras-chave) temáticos no âmbito do assunto, bem como ter requerido a ponderação sobre a pertinência das relações que se podem estabelecer no espaço de campos semânticos e lexicais, levou-nos naturalmente à consciencialização de alguns dos problemas inerentes ao desenvolvimento de sistemas automáticos de recolha de informação. Confrontados com o ato de selecionar informação relevante, leva-nos a crer que a inteligência necessária para dar respostas a questões de natureza complexa é um desafio enorme mas essencial no desenvolvimento dos sistemas automáticos para encontrar respostas não triviais. Acrescenta-se que o conhecimento prévio do assunto tornou a pesquisa, por vezes, mais facilitada e eficiente, revelando que a operação de procura está dependente do aporte de erudição de quem a executa. De facto, se em algumas questões se revelou uma mais-valia a cultura geral dos elementos do grupo (know-how sobre futebol foi utilizado em tópicos relacionados com o desporto; conhecimentos de artes foram utilizados em tópicos relacionados com a música), a par da entreajuda que se fomentou entre todos os elementos, a experiência que o grupo já detém na formulação de termos para pesquisa de informação permitiu um maior ajuste das palavras-chave a submeter ao motor de busca.

4 O Resultado

Mais do que destacar o resultado obtido pelo *LudIT* no Págico, gostaríamos de observar que o sucesso da classificação alcançada foi a consequência do empenho do grupo, constituído por 6 elementos motivados pelo desafio, os quais, por serem investigadores, estão naturalmente treinados para compreender a indispensabilidade de aferir a pertinência quando se pesquisam dados e se testam práticas. A busca de informação, para ser pertinente, deve ser muitas das vezes efetuada através de temas - focos que, numa primeira observação, não estão diretamente relacionados com o assunto. Esta tarefa torna-se seguramente de mais difícil execução se efetuada por meios automáticos. Na verdade, o facto de o *LudIT* ter saído tão bem-sucedido do desafio lançado mostra, em nosso entender, que existe ainda um fosso significativo entre o desempenho humano e o desempenho automático na obtenção de respostas que requerem uma interpretação mais fina em termos de relações semânticas, lexicais e pragmáticas. O resultado mostra igualmente que foi feito um esforço, quer temporal quer de representatividade, ao se ter tido como um objetivo interno responder de forma tão completa quanto possível a todas as questões levantadas pelo desafio.

5 A Conclusão

Vivemos numa sociedade de informação com necessidade de eficiência. Toda a tecnologia que nos envolve tem sido desencadeada por esta urgência de sistemas de busca eficaz. As necessidades de informação são cada vez mais complexas. Desenvolver sistemas automáticos capazes de encontrar respostas a perguntas complexas, em língua portuguesa, é um desafio tão interessante quanto pertinente.

A participação humana num desafio como o definido pelo Págico - Português Mágico mostrou-se interessante e também cativante, uma vez que foi capaz de induzir a necessidade de dar respostas de forma completa. O resultado dessa participação humana pode ser uma mais-valia para validar ou comparar sistemas automáticos. Pode servir também para detetar debilidades de abrangência da Wikipédia.

Tomando a ideia deste desafio, talvez seja possível definir, num futuro próximo, outros desafios, alargados a públicos mais vastos, seguindo a ideia de colaboração on-line para solucionar problemas reais.

Referências

Costa, Luís, Cristina Mota, e Diana Santos. 2012. SIGA, a Management System to Support the Organization of Information Retrieval Evaluations. Em Helena Caseli, Aline Villavicêncio, António Teixeira, e Fernando Perdigão, editores, *Computational Processing of the Portuguese Language, PROPOR'2012*, pp. 284–290, Berlim/Heidelberg. Springer.

Wikipédia. 2012. Wikipédia: A enciclopédia livre, Abril, 2012. `http://pt.wikipedia.org/`.