Uma Abordagem ao PÁGICO baseada no Processamento e Análise de Sintagmas dos Tópicos

Ricardo Rodrigues
CISUC, Universidade de Coimbra
rmanuel@dei.uc.pt

Hugo Gonçalo Oliveira
CISUC, Universidade de Coimbra
hroliv@dei.uc.pt

Paulo Gomes
CISUC, Universidade de Coimbra
pgomes@dei.uc.pt

Resumo

Este artigo descreve a abordagem ao PÁGICO seguida pelo sistema RAPPORTÁGICO. Trata-se de uma abordagem centrada na indexação dos artigos da Wikipédia, na identificação de sintagmas nas frases dos tópicos dados, e no seu posterior processamento e análise, de forma a facilitar a correspondência entre tópicos e artigos que lhes possam servir de resposta. Os sintagmas facilitam a identificação de pequenas estruturas com diferentes papéis dentro da frase. Antes de serem utilizados para consulta, alguns sintagmas sofrem manipulações, como, por exemplo, a expansão das palavras que os constituem em palavras de significado semelhante (sinónimos). Embora haja ainda um longo caminho a percorrer, o sucesso da abordagem traduziu-se, em termos de resultados, na obtenção de uma pontuação com algum destaque entre todas as participações no PÁGICO, especialmente naquelas automáticas.

Palavras chave

Págico, Rapportágico, Rapport, Onto.PT, Sinónimos, Desambiguação do Sentido das palavras, Wikipédia, Análise de Sintagmas

1 Introdução

Citando a própria organização, "o PÁGICO é uma avaliação conjunta na área de recolha de informação em português, que tem por objectivo avaliar sistemas que encontrem respostas não triviais a necessidades de informação complexas, em língua portuguesa" (Santos, 2012). Na prática, o PÁGICO traduziu-se numa tarefa de recolha de informação sobre parte da versão portuguesa da Wikipédia, e é neste contexto que foi desenvolvida e aplicada a abordagem do RAPPORTÁGICO. Procura-se, neste artigo, descrever os vários passos seguidos por esta abordagem à tarefa proposta no PÁGICO.

Esta abordagem teve como ponto de partida a conjugação de esforços dos trabalhos de doutoramento dos dois primeiros autores:

- O projecto RAPPORT, que aborda a resposta automática a perguntas para o português. Deste projecto foi utilizada a análise gramatical feita a textos — neste caso, aos tópicos, que podem ser considerados como perguntas — extraindo-se e identificando-se, em última instância, os sintagmas de cada frase.

- O projecto ONTO.PT (Gonçalo Oliveira e Gomes, 2011b; Gonçalo Oliveira e Gomes, 2012), que tem como objectivo a criação de uma ontologia lexical, estruturada de forma semelhante a uma *wordnet*, também para o português. Deste projecto utilizou-se a base de sinónimos, que permitiu alargar o número de correspondências entre frases nos artigos e as frases dos tópicos. A estrutura do ONTO.PT foi também utilizada para realizar a desambiguação do sentido das palavras que foram expandidas em sinónimos.

O restante documento divide-se pelas seguintes secções: descrição da abordagem e das várias partes que a constituem (Secção 2); uma secção mais focada no processamento dos tópicos e sua transformação em consultas (Secção 3); caracterização das várias corridas submetidas a avaliação (Secção 4); análise e discussão dos resultados (Secção 5); e, finalmente, as conclusões que foram possíveis obter tanto dos resultados em si, como da reflexão *a posteriori* sobre os vários aspectos da abordagem, identificando os pontos fortes e os pontos fracos da mesma, e ainda algumas ideias para trabalho futuro (Secção 6).

2 Abordagem

Como já referido, o RAPPORTÁGICO surge da combinação de alguns elementos dos trabalhos de doutoramento dos autores, nomeadamente ao nível da análise sintáctica de textos e da identificação de sinónimos de palavras em contexto, tirando partido da estrutura de uma ontologia lexical. A abordagem propriamente dita pode ser dividida em quatro partes:

1. **Indexação** dos conteúdos dos artigos;
2. **Análise** e **processamento** das frases dos tópicos, que podem ser vistas como perguntas, com foco nos **sintagmas** que as constituem;
3. **Pesquisa** no índice de conteúdos, utilizando consultas (em inglês, *queries*) geradas no passo anterior, e identificação dos artigos correspondentes às respostas;
4. **Tratamento** das respostas.

O primeiro passo consiste na indexação de todos os artigos da versão portuguesa da Wikipédia presentes na *colecção* produzida para o PÁGICO (Simões, Mota e Costa, 2012). Para o efeito, optou-se pela utilização do motor de pesquisa Lucene (Hatcher e Gospodnetic, 2004), que permitiu criar um índice de documentos com dois campos, nomeadamente o endereço e o conteúdo do artigo. No entanto, com vista à optimização da pesquisa, apenas o conteúdo do artigo foi efectivamente indexado. A utilização do Lucene traz consigo, essencialmente, duas vantagens:

- Facilita a pesquisa de documentos (neste caso, artigos) que correspondam às consultas feitas através de texto, e fá-la de forma célere;
- Permite que, ao processar os conteúdos dos artigos, as palavras sejam normalizadas. No nosso caso, utilizou-se o analisador *portuguese analyzer* (disponibilizado nas contribuições do Lucene), para obter os radicais (*stemming*) das palavras, o que permite, posteriormente, uma comparação mais abrangente entre *queries* e entradas do índice.

Ao realizar o *stemming* são ignoradas, por exemplo, formas e conjugações verbais, bem como números e géneros de nomes e adjectivos. Por exemplo, as conjugações `vence`, `venceram` e `venceremos` são todas normalizadas como `venc`. Isto, à partida, aumentará o número de correspondências entre as *queries* e os conteúdos dos artigos, nem que seja pelo facto de, ao nível das formas verbais, existirem tempos diferentes entre as constantes nos tópicos e aquelas nos artigos — redução de verbos. O mesmo se poderia dizer em termos de número nos nomes — redução de plurais. Isto sem mencionar ainda outras reduções possíveis sobre os conteúdos dos artigos e dos tópicos, como se podem encontrar em Orengo e Santos (2007). Tal levou-nos a descartar, logo de início, uma abordagem sem utilização de um radicalizador.

Apesar da utilização do *stemming* trazer vantagens notórias, traz consigo também algumas desvantagens. Para além de aumentar um pouco a ambiguidade, a principal limitação do *stemming* é o facto de tratar todas as palavras de forma idêntica, independentemente da sua função na frase. Para evitar este problema, havia inicialmente a intenção de normalizar as palavras através da sua lematização, com recurso a um lematizador criado pelo primeiro autor. Contudo, o processo de lematização sobre a *colecção* da Wikipédia revelou-se extremamente demorado, na ordem dos vários dias, o que levou ao seu abandono, por falta de tempo, e à adopção do método de *stemming* fornecido de raiz pelo Lucene — o já referido *portuguese analyzer*.

No segundo passo da abordagem, as frases dos tópicos passam por vários tipos de processamento, com a finalidade de construir uma consulta já preparada para interrogar o índice criado pelo Lucene. Por se tratar do passo mais complexo da nossa abordagem, reservou-se a Secção 3 para descrição dos vários níveis de processamento sofridos pelas frases de cada tópico.

Dada uma consulta, o terceiro passo consiste apenas na utilização do Lucene para obter os artigos mais relevantes. Cada fase de processamento pode gerar uma ou mais restrições que os artigos pesquisados devem respeitar. As restrições são concatenadas na consulta, usando o operador `AND` (disponibilizado pelo Lucene para utilização em *queries*). Também se definiu que só seriam tomados em conta os resultados do Lucene com pontuação superior a zero, ordenados de acordo com a sua relevância para a pesquisa, sendo também ignorados aqueles que se encontrassem fora dos primeiros n devolvidos. No caso da participação oficial no PÁGICO, definimos empiricamente $n = 25$, para todos os tópicos.

Após receber o conjunto de artigos considerados relevantes para a consulta, falta apenas um último passo. Aí, à partida, eliminam-se automaticamente, do conjunto anterior, artigos de tipos que sabemos de antemão não se tratarem de eventuais respostas, tais como: páginas re-

lacionadas com a estrutura da Wikipédia (e.g., páginas começadas por Wikipédia, Portal, Lista ou Anexo); páginas de desambiguação; artigos começados por dígitos; artigos referentes a disciplinas (e.g., Economia, Historiografia, Demografia, etc.); páginas começadas com palavras com o sufixo "ismo" (e.g., Anarquismo, Academicismo, Abolicionismo, etc.). Note-se que a aplicação da lista de exclusões apenas é efectuada após a remoção dos resultados fora dos 25 primeiro devolvidos — uma opção que muito provavelmente agora seria diferente, alterando-se a ordem destes dois passos.

Relativamente à lista de resultados que devem ser excluídos, alguns dos seus elementos têm como evidente a sua inclusão nessa lista, especificamente aqueles relativos à própria estrutura da Wikipédia; já outros foram incluídos com base na análise do tipo de respostas pretendidas e na análise de resultados que eram recorrentes, mas que nunca conteriam a resposta pretendida. Por exemplo, verificámos que as respostas esperadas eram sempre casos concretos e não abstracções, como disciplinas, movimentos, princípios ideológicos, etc.

3 Processamento dos tópicos

Esta secção é dedicada às etapas de processamento das frases nos tópicos do PÁGICO. Como referido na Secção 2, esta é a fase mais complexa da abordagem seguida. O seu objectivo é analisar e processar a frase de cada tópico de forma a construir a consulta que será feita ao Lucene. Cada etapa de processamento é opcional e pode dar origem a uma ou mais restrições que são adicionadas à consulta. Apresentam-se aqui as quatro etapas, nomeadamente a identificação dos sintagmas, a identificação da categoria da resposta, a expansão de sinónimos e ainda a expansão de país ou nacionalidade.

3.1 Identificação de sintagmas

A opção pela identificação dos sintagmas nas perguntas tem por base a convicção de que será mais vantajoso manipular, numa frase, as palavras em grupos, onde estas possam ter algum tipo de relação entre si, em oposição a considerar uma frase como um mero conjunto de palavras sem qualquer relação aparente (à excepção de pertencerem à mesma frase e seguirem determinadas regras gramaticais).

Há, para tal, uma solução de certa forma evidente: a utilização dos sintagmas nominais (SNs) e dos sintagmas verbais (SVs) que constituem os tópicos. Com base na identificação de sintagmas, definou-se uma heurística, que funciona para a maior parte dos casos a apresentados, e que ajudou a reconhecer os elementos mais importantes do tópico: o primeiro SN — mais especificamente, o nome(s) nele presente(s) — será o alvo ou categoria do tópico, enquanto que o(s) SV(s), bem como os restantes SNs, permitem identificar restrições sobre a categoria.

Para se chegar aos sintagmas das frases de cada um dos tópicos, realizaram-se os dois passos seguintes:

- Etiquetagem da categoria gramatical (em inglês, *POS tagging*), com base no analisador (*POS tagger*) do projecto OpenNLP[1], e na utilização dos modelos treinados para a língua portuguesa, também disponibilizados pelo mesmo projecto. Veja-se um exemplo da anotação produzida numa das frases usadas no PÁGICO:
  - Frase original:
    `Filmes sobre a ditadura ou sobre o golpe militar no Brasil`
  - Com etiquetagem gramatical:
    `Filmes\N sobre\PRP a\ART ditadura\N ou\CONJ-C golpe\N militar\ADJ em\PRP o\ART Brasil\PROP`
- Identificação de sintagmas (em inglês, *chunking*), onde se aplicou um conjunto de regras para agrupamento de palavras baseadas na sua etiqueta gramatical. Após a identificação de sintagmas, a frase anterior daria origem às seguintes estruturas:
  - Com identificação de sintagmas:
    `{Filmes}\SN sobre\SP {a ditadura}\SN ou sobre\SP {o golpe militar}\SN em\SP {o Brasil}\SN.`

Sobre a etiquetagem gramatical, interessa referir que houve alguns cuidados na utilização do *POS tagger* do projecto OpenNLP. Por exemplo, procurou-se garantir que nomes compostos (e.g., nomes de pessoas, países, locais) fossem agregados e identificados como um único elemento por parte do *POS tagger*, de forma a facilitar a identificação e posterior análise e manipulação dos SNs. Para o efeito, e na prática, os termos das frases foram processados previamente, nomeadamente através do reconhecimento de entidades mencionadas (Santos e Cardoso, 2007; Mota e Santos, 2008), ignorando-se a eventual classificação, uma vez que apenas era importante, no caso de nomes compostos,

[1] `http://incubator.apache.org/opennlp/`

saber que estes estavam agregados. Por exemplo, pretendia-se que "Universidade de Coimbra" fosse classificada ao nível da etiquetagem gramatical como {Universidade de Coimbra}\N e não como Universidade\N de\PRP Coimbra\N.

Relativamente à identificação de sintagmas, note-se que esta não é, de forma alguma, perfeita; contudo, a identificação que faz dos SNs e dos SVs (centrada mais na identificação de nomes e artigos, num caso, e de formas verbais simples ou compostas, no outro) é, à partida, suficiente para os propósitos da abordagem. As regras utilizadas para identificação de sintagmas foram extraídas do recurso Bosque (Freitas, Rocha e Bick, 2000), disponibilizado pela Linguateca[2], tendo sido feita uma análise da frequência com que etiquetas gramaticais são agrupadas num mesmo sintagma. Após a divisão das perguntas em sintagmas, estes passam a ser o principal elemento no processamento das perguntas.

3.2 Categoria da resposta

Considera-se que o primeiro nome do primeiro SN de cada tópico é o alvo do tópico, ou seja, este nome é uma categoria a que todas as eventuais repostas têm de obedecer. Por outras palavras, esse nome pode ser considerado como um hiperónimo das entidades que serão dadas como resposta, um pouco à semelhança do que fazem Ferreira, Teixeira e da Silva Cunha (2008) para identificar a categoria de entidades mencionadas, que consideram também que a primeira frase num artigo da Wikipédia define normalmente a entidade a que o artigo se refere.

Apesar de em *corpora* existirem vários padrões textuais que indicam a relação de hiperonímia, quando o texto consiste em definições, o padrão <hipónimo> é um <hipernónimo> (*is a* em inglês) sobressai. Isto acontece porque uma forma comum de definir um conceito é através da estrutura: género próximo (*genus*), que é normalmente um hiperónimo, e diferença (*differentia*). É desta forma, aliás, que muitas definições de dicionário são estruturadas (veja-se, por exemplo, Amsler (1981)). Vejam-se também os trabalhos de Snow, Jurafsky e Ng (2005) ou Navigli e Velardi (2010), para o inglês, e Freitas et al. (2010), para o português, onde este padrão é utilizado. No contexto da Wikipédia, o padrão é um já se mostrou também produtivo na aquisição de hiperonímia, como é o caso dos trabalhos de Herbelot e Copestake (2006), para o inglês, e Gonçalo Oliveira, Costa e Gomes (2010), para o português.

Sendo assim, na construção da pesquisa a realizar, começa-se por colocar o padrão anterior antes da categoria. Assim, por exemplo, se a categoria for filme (o nome no primeiro SN), a primeira parte da pesquisa será (é um filme) OR (são um filme) OR (foi um filme) OR (foram um filme). Note-se que não houve preocupação em fazer a concordância em número porque, após o *stemming*, esta acabaria por ser ignorada.

3.3 Expansão de sinónimos

De forma a aumentar a abrangência da pesquisa, no RAPPORTÁGICO é possível indicar alternativas a algumas palavras. Neste caso, as alternativas serão palavras com o mesmo significado, ou seja, sinónimos. Para indicar essas alternativas na consulta, é utilizado o operador OR. Apesar de ser possível, por exemplo, obter sinónimos de qualquer palavra de categoria aberta, apenas realizámos experiências onde obtivemos sinónimos do nome que representa a categoria e ainda dos SVs constituídos por apenas um verbo.

Por exemplo, a categoria músico, pode ter como alternativas as palavras musicista ou instrumentista que, em alguns contextos, têm o mesmo significado. Da mesma forma, os verbos escrever e utilizar podem ter como alternativas, respectivamente, as palavras redigir e grafar, e as palavras usar e empregar.

Após se ter verificado que a expansão dos sinónimos da categoria aumentava a dispersão de respostas, na nossa participação oficial no PÁGICO, limitámo-nos a obter os sinónimos de verbos (ver Secção 4). Acabámos, no entanto, por enviar duas corridas não oficiais com expansão de sinónimos de categorias.

Como base de sinónimos, foram utilizados os *synsets* do ONTO.PT, uma nova ontologia lexical para o português, construída automaticamente a partir de recursos lexicais, e estruturada de forma semelhante à WordNet de Princeton (Fellbaum, 1998). No contexto da WordNet, *synsets* são grupos de palavras sinónimas, que podem ser vistos como a lexicalização de conceitos da linguagem natural. Idealmente, uma palavra pertencerá a um *synset* por cada um dos seus sentidos, e palavras que, em determinado contexto, possam ter o mesmo significado, deverão estar incluídas em, pelo menos, um mesmo *synset*.

Na versão utilizada do ONTO.PT, os *synsets* existentes consistiam nos *synsets* de um *thesaurus* electrónico da língua portuguesa, criado manualmente, o TeP (Maziero et al., 2008). Antes de ser utilizado, o TeP foi enriquecido automaticamente (Gonçalo Oliveira e Gomes, 2011a) com

[2] http://www.linguateca.pt

informação de sinonímia na rede léxico-semântica CARTÃO (Gonçalo Oliveira et al., 2011) que, por sua vez, foi extraída a partir de três dicionários electrónicos do português.

Como palavras com mais de um sentido podem estar incluídas em mais de um *synset*, a obtenção de sinónimos não é trivial, e implica que seja feita a correspondência entre a ocorrência da palavra e o seu sentido mais próximo. Para tal, é necessário utilizar um algoritmo para desambiguar o sentido das palavras (veja-se Navigli (2009) para uma revisão de técnicas para esta tarefa), através da selecção do *synset* correspondente ao significado da palavra no contexto do tópico. Foram utilizados dois algoritmos de desambiguação diferentes, ambos baseados na exploração da estrutura do ONTO.PT, ou seja, nos *synsets* e nas relações entre estes.

Os dois métodos partem do contexto $P = \{p_1, p_2, ..., p_n\}$, e de um conjunto de *synsets* candidatos $C = \{S_1, S_2, ..., S_m\}$. O contexto P inclui, neste caso, todos os nomes e verbos na descrição do tópico. Como as palavras nos *synsets* se encontram lematizadas, nesta fase, as palavras do contexto são também elas alvo de lematização. Todos os *synsets* que incluem a categoria são candidatos e por isso fazem parte de C. Cada um dos dois métodos, descritos de seguida, varia na forma em que é escolhido o *synset* mais adequado, dentro dos candidatos:

- **Bag-of-Words**: para cada candidato, é construído um conjunto $R = \{q_1, q_2, ..., q_p\}$ que inclui as palavras de todos os *synsets* que, no ONTO.PT, se relacionam com o *synset* em questão. O *synset* escolhido é aquele que maximiza a semelhança com o contexto, calculada através da aplicação do coeficiente de *Jaccard*, uma medida bastante comum para esta tarefa:

$$Jaccard(P, R) = \frac{|P \cap R|}{|P \cup R|}$$

  Este método de desambiguação acaba por ser uma a adaptação do algoritmo de Lesk (Lesk, 1986), com duas pequenas diferenças. Primeiro, no algoritmo de Lesk o "contexto" do sentido constrói-se não só com as palavras do *synset*, mas também com as palavras na definição e em frases exemplo. No entanto, como no ONTO.PT essa informação não existe, utilizamos todas as palavras em *synsets* relacionados. Além disso, existe uma diferença na forma de calcular a *sobreposição*. Enquanto que, no algoritmo de Lesk, apenas é utilizado o número de termos comuns, na nossa abordagem é utilizado o coeficiente de *Jaccard*. Ainda que esta opção deva ser futuramente avaliada, a nossa escolha recaiu sobre este coeficiente para que não houvesse um enviesamento na escolha de *synsets* com maiores "contextos", já que, utilizando a medida original, estes teriam maior probabilidade de ter mais palavras em comum com o contexto do tópico.

- **Personalized PageRank**: o método PageRank (Brin e Page, 1998) é normalmente utilizado para ordenar os nós de um grafo de acordo com a sua importância. Foi, no entanto, já utilizado para resolver vários problemas, incluindo a desambiguação de palavras com base na WordNet (Agirre e Soroa, 2009). A nossa implementação é baseada no último trabalho, e utiliza todo o ONTO.PT. Para tal, considera-se que o ONTO.PT é um grafo $G = (N, A)$ com $|N|$ nós, que representam os *synsets*, e $|A|$ arcos sem orientação, para cada relação entre dois *synsets*. Insere-se depois em G um novo nó para cada palavra p_i no contexto. Essas palavras são ligadas a todos os *synsets* que as incluem, desta vez através de um arco direccionado. Se os pesos iniciais forem distribuídos uniformemente apenas aos nós inseridos, é de esperar que, após algumas interacções, o PageRank tenha atribuído maior peso aos *synsets* mais relevantes, dado o contexto.

Para impedir que, quando são seleccionados *synsets* com muitos elementos, a consulta tome proporções demasiado grandes e inclua palavras pouco frequentes, apenas se utilizam como alternativas sinónimos com mais de vinte ocorrências nos *corpora* do serviço AC/DC (Santos e Bick, 2000). Para tal, foram consultadas as listas de frequências disponibilizadas pela Linguateca[3].

3.4 Expansão de nacionalidade ou de país

Sabendo de antemão que os tópicos do PÁGICO se iriam concentrar na cultura lusófona, foi incluída uma fase adicional no processamento dos tópicos, especialmente dedicada a optimizar a expansão de expressões relacionadas com os oito países lusófonos e respectivas nacionalidades. Esta fase subdivide-se em duas partes:

- Para cada ocorrência de uma nacionalidade dos países lusófonos, inclui-se na consulta, como alternativa, o nome do país. Por exemplo, o processamento do sintagma

[3] `http://www.linguateca.pt/ACDC/`

`futebol brasileiro`, dá origem às alternativas `(futebol brasileiro) OR (futebol AND Brasil)`.

- A cada ocorrência de expressões como `país lusófono`, `língua portuguesa`, ou `antiga colónia` foi dada como alternativa o nome de cada um dos países lusófonos. Assim, por exemplo, ao processar o sintagma `país lusófono`, obtém-se a seguinte restrição: `(país lusófono) OR Portugal OR Brasil OR Angola OR Moçambique OR (Cabo Verde) OR (Guiné Bissau) OR (São Tomé e Príncipe) OR Timor`.

Procurou-de assim, e neste caso, tornar as consultas relacionadas com este aspecto tão abrangentes quanto possível sem, no entanto, levar a uma perda de precisão das mesmas.

4 Breve descrição das corridas

A participação oficial do RAPPORTÁGICO no PÁGICO foi constituída por três corridas. Em comum, todas as corridas fazem a identificação dos sintagmas e utilizam cada SN e SV, quando presentes, como restrição; para todas as corridas é identificada a categoria e utilizado o padrão `é um`; e em todas é feita a expansão de país e nacionalidades.

As diferenças entre cada corrida são as que se seguem:

1. A primeira, que pode ser vista como uma *baseline* ao nível da expansão de sinónimos, não tem nada a mais para além dos aspectos acabados de identificar;
2. A segunda faz expansão de sinónimos dos sintagmas verbais com apenas um verbo, utilizando o método *Bag of Words* na desambiguação de termos;
3. A terceira é idêntica à segunda, mas utiliza o método *Personalized PageRank* na desambiguação de termos.

Além das três corridas oficiais, foram enviadas mais duas corridas fora do período oficial. Nas duas corridas adicionais, além da expansão de SVs, é feita a expansão da categoria (o nome no primeiro SN) em sinónimos. Cada uma dessas duas corridas utiliza também um dos dois métodos de desambiguação (à semelhança da segunda e da terceira corrida).

A título de curiosidade, para cada verbo que sofreu a expansão de sinónimos, foram obtidos, em média, 11,6 e 6,5 sinónimos na segunda e na terceira corrida, respectivamente. Já relativamente à expansão das categorias, foram obtidos, em média, 5,9 e 6,4 sinónimos para cada categoria, respectivamente na quarta e na quinta corrida — as corridas extra-oficiais.

É de referir que, dias antes de terminarmos a escrita deste artigo, verificámos a existência de problemas no código da desambiguação, que estavam a impedir que o contexto fosse tomado em conta. Desta forma, nas cinco corridas aqui descritas, a escolha do melhor *synset* foi, na realidade, feita da seguinte forma: no algoritmo *Bag-of-Words*, estaria a ser escolhido um *synset* aleatório, enquanto que nas restantes corridas estava a ser aplicado um *PageRank* simples, e não o *Personalized PageRank*. Ou seja, era sempre escolhido o *synset* que, dada a estrutura do grafo e sem qualquer contexto, tivesse melhor pontuação. Apesar de tudo, principalmente em palavras com pouca ambiguidade, esta situação não terá afectado em demasia os resultados, mas contamos fazer essa avaliação brevemente, de forma semelhante à avaliação das restantes corridas não oficiais.

5 Resultados

Notem-se os resultados oficiais comparados da nossa abordagem: de um total de 12 submissões repartidas pelos vários participantes, o RAPPORTÁGICO obteve o quinto, o sexto e o sétimo lugares, com pontuações de 25,0081, 23,7379, e 19,0693 pontos, para as corridas 3, 2 e 1, respectivamente.

Pode-se observar na Tabela 1 uma súmula dos resultados da abordagem proposta, onde são apresentados o número total de respostas certas, bem como o número de tópicos que obtiveram pelo menos uma resposta certa, para cada uma das corridas, com diferentes pontos de corte. São apresentados também os resultados para diversos pontos de corte (limites) relativamente ao número de respostas submetidas por tópico, a precisão, a pseudo-abrangência (pseudo, na tabela) e a pontuação correspondente — relembramos que os resultados oficiais se referem a um ponte de corte correspondente a um máximo de 25 respostas por tópico, como já referido anteriormente, sendo identificados a carregado; a itálico encontram-se os melhores resultados parcelares para cada uma das corridas, quando aplicável.

É possível observar a existência de uma certa proporcionalidade nos resultados dos diversos pontos de corte, já que tanto aumentam o número de respostas submetidas, como o número de res-

Corrida	Limite	# Respostas	# Submetidas	Precisão	Pseudo	Pontuação	# Tópicos
	5	86	512	*0,1680*	0,0383	14,4453	47
	10	122	918	0,1329	0,0543	16,2135	51
1	15	147	1275	0,1153	0,0654	16,9482	54
	20	164	1577	0,1040	0,0730	17,0551	56
	25	***181***	**1718**	**0,1054**	***0,0805***	***19,0693***	***59***
	5	90	516	*0,1744*	0,0400	15,6977	50
	10	132	927	0,1424	0,0587	18,7961	53
2	15	164	1289	0,1272	0,0730	18,3986	58
	20	184	1591	0,1157	0,0819	21,2797	58
	25	***203***	**1736**	**0,1169**	***0,0903***	***23,7379***	***59***
	5	92	518	*0,1776*	0,0409	16,3398	48
	10	135	940	0,1436	0,0601	19,3883	53
3	15	166	1305	0,1272	0,0738	21,1157	57
	20	188	1601	0,1174	0,0836	22,0762	58
	25	***208***	**1730**	**0,1202**	***0,0925***	***25,0081***	***59***

Tabela 1: Resultados das Várias Corridas

postas certas e os tópicos com pelo menos uma resposta certa. O mesmo acontece com a pontuação para cada uma das alternativas dos limites. Talvez isso possa ser um indicador de que o ponto de corte inicialmente estipulado pudesse ser ligeiramente mais elevado — contudo, esta análise apenas surgiu *a posteriori*, e quando participámos ainda não estávamos certos de como a abordagem seria avaliada. Note-se também que para os pontos de corte 5, 10 e 15, apesar de haver menos respostas correctas, a precisão é superior àquelas das corridas oficiais, dado o *ratio* mais favorável entre o número de respostas correctas e o número total de respostas submetidas. Já a pseudo-abrangência vai crescendo com o aumento do valor dos pontos de corte.

Quanto às diferenças em termos dos próprios resultados de cada uma das três corridas, apesar de não ter sido possível uma comparação exaustiva das respostas presentes ou ausentes em cada uma das corridas e compará-las com as restantes, através de uma simples análise de linhas de respostas diferentes, foi possível verificar que as corridas mais diferentes em termos de resultados foram a segunda e a terceira, com 123 respostas diferentes, sendo que as respostas diferentes entre a primeira e a segunda foram 78, e entre a primeira e a terceira foram 101. Apesar de as linhas diferentes apenas conterem uma pequena parte de respostas correctas, isto leva-nos a crer que cada uma das corridas obtém um conjunto pequeno de respostas que não são partilhadas.

Há, contudo, um aspecto curioso que deve ser apontado: qualquer uma das corridas conseguiu apresentar respostas correctas para o mesmo número de tópicos. Isto indicará que as várias corridas diferiram essencialmente no número de respostas correctas que apresentaram para cada tópico. Uma hipótese é que os termos constantes da pergunta (ou tópico) *inicial* são os que melhor definem as respostas pretendidas. Todo o restante processamento dos tópicos ajuda essencialmente a encontrar mais alternativas (tanto correctas como incorrectas) de respostas.

Outro aspecto interessante é o facto de as restrições dos tópicos muitas vezes se encontrarem distribuídas pelos conteúdos dos artigos, por várias frases, ou até mesmo parágrafos, o que leva a crer que todo um texto tem importância para a obtenção de respostas a perguntas mais complexas — contrariando a ideia, por vezes recorrente, que o elemento mais importante na obtenção de uma resposta é a descoberta de uma frase específica (com ou sem variações).

Relativamente às duas corridas não oficiais, obtivemos um resultado de certa forma surpreendente: ao contrário do que seria esperado, o número de respostas geradas tinha aparentemente diminuído (1529 e 1519, respectivamente), e também o número de perguntas com resposta tinha diminuído (para 49 e 56, respectivamente).

Após análise, foi possível concluir que o número de respostas geradas nestas duas corridas não tinha diminuído; contudo, muitas das (novas) respostas obtidas, com expansão dos SNs, vieram posteriormente a ser ignoradas por se encontrarem na lista de exclusões — e pelo facto de esta lista só se aplicar após obtenção das respostas através do Lucene. Isto leva-nos a crer que tanto a expansão de SNs e SVs contribuem para uma maior abrangência das *queries*, mas, apesar de tudo, os SVs expandidos mostram-se mais próximos do significado inicial que os SNs expandidos.

Quanto à pontuação destas duas últimas corridas, reflectindo os números das respostas, a quarta corrida obteve 16,1210 pontos, e a quinta 19,7031 pontos. O mesmo aconteceu com a

precisão (0,1027 e 0,1139, respectivamente) e a pseudo-abrangência (0,0698 e 0,0769, respectivamente). Dados estes valores, não julgámos pertinente investigar variações com pontos de corte diferentes.

6 Conclusões

Em termos conclusivos, e fazendo alguma reflexão, podemos afirmar que, apesar de haver ainda um longo caminho a percorrer, os resultados obtidos foram interessantes, tanto mais que as perguntas a concurso acabaram por ser bastante distintas daquelas inicialmente apresentadas como exemplos.

Essas perguntas levaram-nos a crer que a análise da estrutura e do tipo de pergunta seriam os pontos mais importantes das mesmas, indo mesmo ao encontro dos trabalhos do primeiro autor. Contudo, as perguntas a concurso, na prática, não o eram, sendo mais próximas de um enunciado com restrições, o que nos obrigou a repensar toda a estratégia para o PÁGICO.

Em todo o caso, e talvez mesmo por essa alteração, acreditamos que os resultados da abordagem até possam vir a revelar-se mais proveitosos e abrangentes que o inicialmente previsto, permitindo aumentar a abrangência dos trabalhos dos autores, aplicando-os num novo cenário.

Algumas ideias para trabalho futuro incluem, por exemplo, a utilização de uma lista de exclusões mais extensa e precisa, bem como a sua aplicação antes de limitar o número de respostas a devolver. Também gostaríamos de explorar a expansão de alternativas para as categorias, não apenas em sinónimos, mas também no seus hipónimos. Este tipo de expansão iria permitir, por exemplo, que para a categoria `músico` fossem dadas alternativas como `pianista`, `flautista` ou `guitarrista`.

Seria também interessante fazer uma análise extensa dos resultados de cada corrida e saber quais as respostas que só são obtidas por cada uma delas, a razão desse facto e verificar se haveria alguma forma de as combinar.

Outro ponto interessante seria estudar qual o melhor n a considerar na selecção das respostas e ver até que ponto será possível tirar partido do Lucene para identificar mesmo um n diferente para cada conjunto de respostas.

Agradecimentos

Gostaríamos de agradecer à organização do PÁGICO, tanto pela ideia da avaliação conjunta em si, que nos levou à aplicação em contexto diferente de parte do trabalho que temos vindo a desenvolver e a uma reflexão sobre o mesmo, como pela disponibilidade e apoio prestado a questões que colocámos durante e após o concurso, incluindo a avaliação das corridas não oficiais, e que se prolongou também na revisão ao artigo.

Hugo Gonçalo Oliveira é apoiado pela bolsa de doutoramento SFRH/DB/44955/2008 da FCT, co-financiada pelo FSE.

Referências

Agirre, Eneko e Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. Em *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'09, pp. 33–41, Stroudsburg, PA, USA. ACL Press.

Amsler, Robert A. 1981. A taxonomy for english nouns and verbs. Em *Proceedings of the 19th annual meeting on Association for Computational Linguistics*, pp. 133–138, Morristown, NJ, USA. ACL Press.

Brin, Sergey e Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107–117.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Ferreira, Liliana, António Teixeira, e João Paulo da Silva Cunha. 2008. REMMA — Reconhecimento de entidades mencionadas do MedAlert. Em Cristina Mota e Diana Santos, editores, *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas*. Linguateca, pp. 213–229.

Freitas, Cláudia, Paulo Rocha, e Eckhard Bick. 2000. Floresta Sintá(c)tica: Bigger, Thicker and Easier. Em *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, PROPOR'2008, pp. 216–219. Springer-Verlag.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, e Violeta Quental. 2010. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. Em *Livro do IX Encontro de Linguística de Corpus*, ELC 2010.

Gonçalo Oliveira, Hugo, Leticia Antón Pérez, Hernani Costa, e Paulo Gomes. 2011. Uma

rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática*, 3(2):23–38, December, 2011.

Gonçalo Oliveira, Hugo, Hernani Costa, e Paulo Gomes. 2010. Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. Em *Actas do II Simpósio de Informática (INFORUM 2010)*, pp. 537–548. Universidade do Minho.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2011a. Automatically enriching a thesaurus with information from dictionaries. Em *Progress in Artificial Intelligence, Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, volume 7026 of *LNCS*, pp. 462–475. Springer, October, 2011.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2011b. Onto.PT: Construção automática de uma ontologia lexical para o português. Em Ana R. Luís, editor, *Estudos de Linguística*, volume 1. Imprensa da Universidade de Coimbra, Coimbra. No prelo.

Gonçalo Oliveira, Hugo e Paulo Gomes. 2012. Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese. Em *Natural Language Processing and Information Systems, Proceedings of 17h NLDB*, LNCS, pp. No prelo, Groningen, The Netherlands. Springer.

Hatcher, Erik e Otis Gospodnetic. 2004. *Lucene in Action*. Manning Publications, December, 2004.

Herbelot, Aurelie e Ann Copestake. 2006. Acquiring ontological relationships from wikipedia using RMRS. Em *Proceedings of the ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*.

Lesk, Michael. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. Em *Proceedings of the 5th Annual International Conference on Systems documentation*, SIGDOC '86, pp. 24–26, New York, NY, USA. ACM.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo, e Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 390–392.

Mota, Cristina e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, December, 2008.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Navigli, Roberto e Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. Em *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1318–1327, Uppsala, Sweden, July, 2010. Association for Computational Linguistics.

Orengo, Viviane Moreira e Diana Santos. 2007. Radicalizadores versus Analisadores Morfológicos: Sobre a participação do Removedor de Sufixos da Língua Portuguesa nas Morfolimpíadas. Em Diana Santos, editor, *Avaliação Conjunta: um novo Paradigma no Processamento Computacional da Língua Portuguesa*. IST Press, Lisboa, Portugal, pp. 91–104.

Santos, Diana. 2012. Porquê o Págico? *Linguamática*, 4(1), Abril, 2012.

Santos, Diana e Eckhard Bick. 2000. Providing Internet Access to Portuguese Corpora: the AC/DC project. Em *Proceedings of the 2nd International Conf. on Language Resources and Evaluation*, LREC'2000, pp. 205–210.

Santos, Diana e Nuno Cardoso, editores. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, November, 2007.

Simões, Alberto, Cristina Mota, e Luís Costa. 2012. A Wikipédia em português no Págico: adaptação e avaliação. *Linguamática*, 4(1), Abril, 2012.

Snow, Rion, Daniel Jurafsky, e Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. Em *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, pp. 1297–1304.