Volume 4, Número 1- Abril 2012

linguaMATICA

ISSN: 1647-0818

UNIVERSIDADE DE VIGO

Universidade do Minho

Volume 4, Número 1 – Abril 2012

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Editorial

É com grande orgulho que observámos o início do quarto ano da Linguamática. Não é fácil garantir uma publicação periódica, quanto mais quando é dedicada a conteúdo científico, e quando somos forçados a publicar sempre nos mesmos sítios bem cotados. Mas só podemos ser uma publicação de referência se conseguirmos publicar, e por isso queremos agradecer a todos os autores que já participaram nesta aventura, e a todos os membros da comissão científica que sempre nos ajudaram a definir o melhor conteúdo.

Este quarto ano começa com a segunda edição especial da Linguamática que inaugura um novo visual. Sabemos que o conteúdo da revista é o importante, mas se nos der gozo lê-la, o conteúdo torna-se ainda mais cativante. Esperamos que gostem do novo estilo visual.

*Esta edição especial resume uma avaliação conjunta que visa a procura de informação, e nomeadamente a procura de informação na Wikipédia Portuguesa – Págico (**P**ortuguês M**ágico**).*

As avaliações (conjuntas) são actividades que nos merecem o maior respeito. Mas porquê o nosso interesse em avaliações conjuntas? Por experiência própria sabemos que para além do enorme esforço ligado ao trabalho de organização, a "simples" participação traz inevitavelmente a certeza da dúvida: Por um lado é o volume de dados que torna o possível, impossível; o calculável, impraticável; e o trivial, incerto! Por outro lado é aquele algoritmo brilhante que acaba por não ser realista, ou aquele resultado tão esperado que não aparece.

Depois de ultrapassada essa fase traumática, verificamos que o nosso modelo até faz sentido; pode não ser o melhor, mais eficiente, e com melhores resultados, mas interessa e surpreende os outros participantes; e com um bocado de trabalho futuro, os tais avanços esperados até serão possíveis, calculáveis e pelo menos sonháveis! E concluímos que será interessantíssimo voltar a testar uma nova versão numa nova avaliação conjunta...

Estamos perante um laboratório de experimentação — e nisso acreditamos.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Introdução à Edição Especial

Págico — Português Mágico

Esta edição pretende apresentar a uma audiência internacional o Págico, uma avaliação conjunta inovadora organizada pela Linguateca. O Págico foi organizado de forma ao seu encontro final ser um encontro satélite do PROPOR 2012, a cuja organização agradecemos o convite.

Dada a pouca vitalidade do mercado editorial português, por um lado, que não consegue publicar contribuições de interesse científico com pouca audiência, e a existência e pujança de uma revista de acesso aberto dedicada aos temas do processamento das línguas ibéricas, incluindo, naturalmente, o português, por outro, a Linguamática configurou-se como a opção de excelência.

Uma avaliação conjunta pressupõe sempre muito trabalho de organização e muitas decisões e escolhas que, para mais tarde serem úteis à comunidade que pretende servir, têm de ser documentadas e fundamentadas.

Por outro lado, é importante compreender a motivação e as estratégias dos participantes, e uma edição sobre uma avaliação conjunta tem como um dos pilares fundamentais a participação. Dar voz e disseminação aos participantes é, pois, uma das obrigações mais gratas dos organizadores, que ao divulgarem o trabalho feito com base na sua organização podem apreciar os frutos desse trabalho.

Além disso, o resultado **conjunto** *da organização e da participação permite construir recursos de treino e de avaliação, que, embora constituam um trabalho hercúleo para a organização, são talvez o que justifica em maior grau a existência de um volume de uma revista dedicada a essa iniciativa, porque construiu algo que ficará – esperamos que por muito tempo – disponível a tantos quantos queiram*

- *investigar a recolha de informação em português*
- *estudar a cultura lusófona*
- *criar sistemas de resposta automática a perguntas, de recolha de informação (RI), de visualização de resultados, etc.*
- *estudar a wikipédia e a sua evolução.*

A presente edição segue o seguinte formato (em que todas as contribuições foram revistas por dois elementos da comissão científica):

1. *vários artigos, escritos por membros da organização do Págico, sobre as opções tomadas e a motivação das mesmas, que podemos considerar pré-avaliação;*
2. *os artigos dos participantes;*
3. *vários artigos, outra vez da organização, descrevendo os resultados do Págico: os problemas e dificuldade da avaliação, a pontuação dos participantes, os recursos finalmente produzidos, e um balanço crítico da iniciativa.*

Embora com esta estrutura tenhamos tentado seguir a evolução temporal da própria avaliação conjunta, e tentássemos que o volume pudesse ser lido do princípio ao fim como um livro, os artigos não foram escritos em sequência. Pelo contrário, encorajámos que fossem mencionadas questões e dados apenas encontrados em artigos "posteriores", quando tal fizesse sentido.

Esperamos que esta edição possa contribuir para um maior conhecimento não só da iniciativa em si mas também dos problemas associados ao processamento da wikipédia em português, à compreensão de utilizadores humanos, e à criação de consultas em RI.

Diana Santos
Cristina Mota
Cláudia Freitas
Luís Costa

Comissão Científica Convidada

Alberto Simões, Universidade do Minho
António Teixeira, Universidade de Aveiro
Belinda Maia, Universidade do Porto
Cláudia Freitas, Pontifícia Universidade Católica do Rio de Janeiro e FCCN/Linguateca
Cristina Mota, FCCN/Linguateca
Diana Santos, Universidade de Oslo e FCCN/Linguateca
Fernando Perdigão, Universidade de Coimbra
José João Dias de Almeida, Universidade do Minho
Luís Costa, FCCN/Linguateca
Luísa Coheur, INESC-ID e Instituto Superior Técnico
Paulo Gomes, Universidade de Coimbra
Sandra Aluísio, Universidade de São Paulo
Stella Tagnin, Universidade de São Paulo
Xavier Gómez Guinovart, Universidade de Vigo

Porquê o Págico? Razões para uma avaliação conjunta

Diana Santos
Linguateca/FCCN & Universidade de Oslo
d.s.m.santos@ilos.uio.no

Resumo

Este artigo apresenta a motivação da avaliação conjunta Págico - Português Mágico, organizada pela Linguateca em 2011-2012 como uma medida para (i) incentivar o desenvolvimento de sistemas de ajuda à procura de informação em português; (ii) avaliar a wikipédia em português; (iii) estudar a interação humana na procura de respostas, e compará-la com as características dos sistemas automáticos. Depois de fazer uma pequena descrição da própria tarefa e de iniciativas relacionadas, mencionando também a organização de anteriores avaliações conjuntas pela Linguateca, cada uma das questões acima mencionadas é descrita e problematizada.

Palavras chave

Avaliação, extração de informação, recolha de informação, resposta a perguntas, português, lusofonia, avaliação conjunta, wikipédia

1 Apresentação

Como descrito no editorial do presente volume, o Págico foi uma avaliação conjunta que decorreu em 2011-2012 sobre a wikipédia em português, preparada pela Linguateca a partir da sua versão de Abril de 2011, usando apenas o texto (e não as páginas completas).

Muito brevemente, e citando o folheto de divulgação do Págico (Págico, 2011) criado na altura da sua disseminação, pretendíamos que sistemas (e pessoas) respondessem a perguntas, ou tópicos, com base na wikipédia em português:

> Exemplos de perguntas associadas à cultura e sociedade lusófona em que pretendemos uma resposta agregada, e justificada, com base na informação da wikipédia:
>
> - Que outros resistentes associados a movimentos de libertação privaram com Amilcar Cabral durante a vida deste?
> - Que cientistas ou avanços da ciência podem ser direta ou indiretamente relacionados com os jesuítas da escola de Coimbra?
> - Que gramáticos brasileiros se pronunciaram sobre a questão da "língua brasileira"?
> - Quais os jogadores de futebol de língua portuguesa que passaram por mais de três países estrangeiros na sua vida profissional?

Cada resposta seria uma página da wikipédia, a que estaria associado um conjunto de páginas adicionais que a justificassem, caso a própria página não contivesse informação suficiente para uma pessoa confirmar que era uma resposta válida.

Os sistemas podiam enviar no máximo cem (100) respostas por pergunta, e podiam enviar 3 corridas diferentes.

Para os resultados, a participação e os recursos criados, veja-se os restantes artigos do presente volume, visto que este artigo tem como único objetivo motivar a própria origanização do Págico.

Assim, tentarei explicar porque é que na Linguateca achámos que seria relevante organizar o Págico nos moldes em que foi organizado, dividindo a argumentação em três partes: avaliação da wikipédia, desenvolvimento de sistemas realistas aplicados a tarefas comuns, e estudo de utilizadores interessados em cultura lusófona.

Embora a secção de motivação do sítio do Págico liste cinco diferentes razões que pensámos pudessem motivar possíveis participantes e explicar o interesse duma iniciativa como a nossa, nomeadamente

- As limitações dos sistemas atuais
- A falta de interesse pela lusofonia
- O enviesamento da wikipédia em português
- A necessidade de juntar esforços para o ensino e o ensino da cultura
- O concurso homem-máquina

neste artigo propomos uma categorização diferente, também um pouco com base naquilo que foi acontecendo ao longo do Págico, e que aumentou a nossa compreensão do processo que despoletámos.

Para um balanço, veja-se Santos et al. (2012), mas no que aqui interessa podemos desde já adiantar que os dois últimos pontos não surtiram qualquer efeito: estamos convencidos de que nem o Págico foi – pelo menos até agora – aproveitado em experiências pedagógicas, nem as pessoas se sentiram especialmente interessadas em concorrer com sistemas automáticos.

Pelo contrário, parece que são os desenvolvedores de sistemas automáticos que estão especialmente interessados em publicitar os seus sistemas como capazes de ter um desempenho próximo ou melhor do que o humano (como aconteceu com a participação do Watson na Jeopardy (Thompson, 2010; Ringel, 2011)), e que o espírito de competição das pessoas, pelo contrário, não é especialmente posto em destaque quando os seus competidores são máquinas. Uma das razões para isto pode naturalmente ser que, como o desempenho dos sistemas concorrentes demonstrou (Mota, 2012), ainda estamos muito longe de uma situação em que estes consigam ombrear com os seres humanos.

Quanto à questão pedagógica no ensino de português como língua estrangeira, estamos conscientes de que a causa da não utilização do Págico foi o curtíssimo prazo que houve entre a divulgação da iniciativa e a sua execução, e não deve portanto a observação anterior ser interpretada como uma crítica aos professores que poderiam fazer uso do material do Págico. De facto, esperamos que possam vir mais tarde a fazê-lo, não em "tempo real", mas em tarefas pensadas e apropriadas aos objetivos do ensino de cada um.[1]

2 Apreciação da wikipédia

Não há dúvida de que a wikipédia é uma das fontes mais consultadas mundialmente e, embora não tenhamos dados para os utilizadores da rede de língua portuguesa apenas, pensamos que tal se verifica também ao comparar páginas ou sítios em língua portuguesa. De qualquer forma, o simples facto de ser de acesso aberto e modificável e melhorável por qualquer pessoa que a consulte faz com que seja um dos maiores recursos da Web 2.0 em português, e por isso digna de estudo e de utilização, quer no desenvolvimento de sistemas para o português, quer na criação de formas alternativas de consultar e reutilizar o conhecimento nela contido.

Como é sabido, cada vez mais a informática está dirigida a estudos de utilização para melhorar os seus produtos e para compreender a forma de interagir do público: a massificação da sociedade do conhecimento faz-se também através do teste e escolha de vários procedimentos alternativos, muitas vezes tendo o utilizador contribuído, sem saber, para a melhoria significativa do sistema que usa.

Por outro lado, existem várias controvérsias sobre se deixar que qualquer pessoa edite as fontes de conhecimento permite melhorar a qualidade de uma enciclopédia ou, pelo contrário, inquina a transmissão de conhecimento, e vários estudos – de índoles e objetivos diferentes – têm alegado enviesamento da Wikipédia

- em relação ao peso relativo dos assuntos: Veale (2007) notou a importância desproporcionada de informação sobre livros de ficção científica e personagens de certo tipo de literatura fantástica na wikipédia;
- em relação a questões políticas: Hagen (2008) defende que um grupo de pessoas com uma agenda política extremista é capaz de controlar e modificar as informações na wikipédia sem que a maioria da população se dê conta.

São, aliás, bem conhecidas as variadas discussões que de vez em quando assolam a wikipédia em relação a artigos em que diferentes autores têm opiniões radicalmente diferentes, e que os editores têm de congelar.

Contudo, e pelo menos que eu saiba, não tem havido muitas avaliações da wikipédia em português no que se refere ao seu conteúdo e abrangência, nem à sua capacidade de satisfazer os utilizadores que nela procuram[2]. Existe sim uma prática extremamente positiva e interessante dos wikipedistas brasileiros que têm organizado concursos de melhoria da wikipédia no que se refere ao material em português, como é disso prova o presente I GP Wikimedia Brasil.[3] De qualquer maneira, quando iniciámos

[1] Agradeço à Cláudia Freitas por chamar a atenção sobre a possibilidade de o meu comentário sobre este assunto ser entendido como uma crítica.

[2] Claro está que a noção de satisfação de um utilizador é complicada e não passa necessariamente pela garantia de qualidade – uma pessoa pode ficar muito satisfeita por ter aprendido muitas coisas que não sabia ... para mais tarde descobrir que não eram verdade – ou, pior ainda, nunca vir a descobri-lo.

[3] Em `http://pt.wikipedia.org/wiki/Wikip\%C3\%A9dia:GP`.

o Págico não estávamos conscientes desta iniciativa, e pensámos que o Págico poderia dar um contributo para o trabalho de avaliação. Pelo menos posso dar voz à satisfação de ver que no Brasil a língua portuguesa ainda é o veículo preferido de comunicação e conhecimento, como aliás foi-nos comentado pela Belinda Maia a propósito das suas aulas de terminologia na Universidade do Porto: não há dificuldade em encontrar texto técnico e científico em português do Brasil, enquanto que em Portugal os textos técnicos são geralmente escritos (ou pelo menos publicados) em inglês.[4] Ainda reforçando a convicção de que o futuro está no Brasil, veja-se as experiências pedagógicas universitárias de melhoria da wikipédia em português relatadas por Neto (2012).

Depois de usarmos instantâneos (progressivamente maiores) da wikipédia em português em várias avaliações conjuntas internacionais, como o QA@CLEF (Giampiccolo et al., 2008) para responder a perguntas em português (ou cuja resposta estivesse em português), e o GikiP (Santos et al., 2009) e o GikiCLEF (Santos et al., 2010) para responder a um tipo especial e mais complexo de perguntas cuja resposta também se podia encontrar na wikipédia em português, pensámos ter chegado a altura de dedicar uma avaliação conjunta apenas a responder a perguntas em português cuja resposta estivesse em português.

Isto não só porque a Linguateca se dedica ao processamento computacional do português, mas porque a parcela de atenção e de cobertura do português por oposição às outras línguas resultou muito diminuta nessas avaliações anteriores, como discutimos em Santos e Cabral (2009).

3 Desenvolvimento de sistemas realistas

Uma das principais razões para a organização de avaliações conjuntas pela Linguateca (Santos, 2007) é a nossa convicção de que essa organização leva a comunidade a desenvolver sistemas e a resolver problemas práticos que resultam no avanço da área como um todo, e que levam ao eventual surgimento de sistemas com aplicação prática.

Embora seja claro que, pelo menos desta vez, tal organização não levou a esse efeito desejado, com apenas dois participantes automáticos que não parecem ter aplicado muito esforço à tarefa oferecida pelo Págico, usando-o apenas tangencialmente como verificação ou teste de partes do seu trabalho, a motivação desta tarefa era clara nesse sentido.[5]

Ainda existe muito pouco apoio para respostas agregadas, exceto no caso da distribuição geográfica, onde os mapas são populares e usados em muitas interfaces. Além disso, encontram-se também em alguns sistemas os cronogramas ou linhas temporais (Heyer, Holz e Teresniak, 2009; Heyer et al., 2011), e, naturalmente, os grafos, que geralmente ligam objetos semelhantes (Dorow, 2006; Widdows, 2004). Mas os grafos, para que a sua visualização seja humanamente viável, pressupõem tentativas de diminuir o número de dimensões, veja-se por exemplo Speer, Havasi e Lieberman (2008).

Contudo, a obtenção de respostas variadas e múltiplas a uma mesma necessidade de informação (o que tecnicamente se pode chamar respostas abertas a um assunto sobre o qual não conhecemos de antemão as respostas) é uma tarefa muito mais complicada, não só de avaliar: quem pergunta, aprende, e essa aprendizagem depende muito do quanto já sabia, o que torna a utilidade de um tal sistema muito diferente conforme o perguntador (Freitas et al., 2012). O que é, aliás, algo já bem conhecido e problematizado na área de recolha de informação, veja-se por exemplo a discussão de Saracevic (1995). Além disso, não é fácil comparar medidas de desempenho de sistemas com objetivos dfiferentes, veja-se Su (1998).

Durante a organização do QA@CLEF, esse foi um assunto muito discutido, mas devido à inércia provocada por um grande número de organizadores, foi apenas atacado em avaliações conjuntas mais pequenas e, infelizmente, com muito menos participação, nomeadamente o Wika (Jijkoun e de Rijke, 2007) e o GikiP (Santos et al., 2009).

Outra linha de discussão, embora ainda não muito seguida, tem a ver com a variedade e diversidade na apresentação de resultados — veja-se as tentativas nesse campo, principalmente em avaliações de imagens (Karlgren, Clough e Gonzalo, 2006) ou de cariz geográfico (Bucher et al., 2005).

Pensámos, de qualquer maneira, que a tarefa do Págico poderia levar à mentalização da

[4] Realce-se a este propósito, embora não vindo diretamente à baila, que a Linguateca disponibilizou há tempos memórias de tradução português-inglês em áreas especializadas, tal como engenharia industrial, arquitetura, gestão, es estudos literários, e que podem ser úteis para mostrar as diferenças entre as línguas nas áreas respetivas, cf. `http://www.linguateca.pt/Repositorio/`.

[5] Mais uma vez, isto não é para ser lido como crítica, dado o pouco tempo entre a divulgação e a participação: é apenas uma constatação.

comunidade informática para a necessidade de desenvolver sistemas que dessem alguma resposta a este tipo de problemas.

4 Estudo de utilizadores interessados em cultura lusófona

Outra área em que precisamos absolutamente de compensar a falta do peso da cultura lusófona na investigação em recolha de informação a nível mundial é a da investigação de assuntos relacionados com a nossa cultura e língua.

De facto, o português, sendo uma das línguas mais faladas no mundo, tem um peso comercial e cultural comparativamente muito reduzido, ou pelo menos anda arredado da atenção de muitos atores no campo da recolha de informação e/ou do processamento computacional da língua. Senão atente-se aos seguintes indicadores

- Muito raramente o português é uma língua de investigação internacional – contam-se pelos dedos as avaliações envolvendo o português, numa altura em que não é exagero dizer que existe uma avaliação (ou "shared task") semana sim semana não;
- A maior parte das grandes iniciativas do Google não contemplam o português (por exemplo, o Google books ngram viewer[6], além de inglês e chinês, só contempla, o francês, o alemão, o espanhol, o russo – e o hebraico!)

A própria União Latina, cujo objetivo é dinamizar e estudar as línguas românicas, não dá muito peso ao português, e o barómetro que propõe[7] confere apenas a ordem dezasseis ao português. Mas, considerando com mais atenção a forma como os valores foram calculados, apercebemo-nos de que o único elemento original é a fórmula: Todos os outros indicadores provêm de organizações diferentes.[8]

Embora o Observatório da Língua Portuguesa[9] nos últimos tempos tenha ganho um dinamismo apreciável (ao contrário do relatado em Santos (2009)), pode constatar-se que os dados que apresenta provêm... exatamente da mesma fonte.

Voltando ao barómetro da União Latina, é muito estranho que seja o francês a língua mais cotada (entre as línguas românicas). Olhando com mais atenção, outra das bases para calcular o peso das línguas é algo chamado "Index translatorum"[10], que – mais uma vez na mesma data – continha a seguinte observação relativa a Portugal: "1989, 1990, 1991, 1992, 1997, 1998, 1999, 2000, 2001, 2002, 2005, 2006, 2007, 2008 and 2009 currently being processed by the INDEX team". Ou, por outras palavras: os dados ainda não se encontram lá.

Quanto ao estudo relacionado com a cultura (outro estudo que podemos consultar do sítio da União Latina, denominado "Línguas e culturas na Web: Estudo 2007"), e que data (pelo título) de há cinco anos (União Latina – Direção de Terminologia e Indústrias da Língua (DTIL), s.d.), as imprecisões e erros são de tal forma gritantes que nos convencem da pouca fiabilidade em relação ao português: desde "amalia rodriguez" a "antónio cavaco silva", passando por "Otelo de Carvalho", "sofia de mello bryner" e "Luis de Camões", até ao facto de que nas "Letras" o valor de Maio de 2008, 3.752.201 é vinte vezes menor do que o de julho de 2005, 65.323.792[11], ao contrário de todas as outras categorias, tudo nos leva a duvidar de que possamos confiar nos resultados.

Seja como for, em Prado, Pimienta e Álvaro Blanco (2009), os autores discutem os variados problemas que tiveram com as mudanças no funcionamento e existência dos motores de procura, e sugerem que uma das grandes vantagens do seu trabalho é ter usado o mesmo método ao longo de mais de uma década, podendo portanto os números serem usados para estudar a evolução da presença das línguas na rede.

Não querendo criticar estas duas instituições – o observatório da Língua Portuguesa e a União Latina – que, pelo menos, lutam contra a corrente, achamos contudo que estes números demonstram bem o quanto ainda é preciso fazer para levantar o português ao nível que merece.

De qualquer maneira, gostávamos de salientar o trabalho de Calvet (Calvet, 2006; Calvet, 2008), que está na origem do dito barómetro, como merecedor de reflexão e do maior respeito,

[6] `http://books.google.com/ngrams`

[7] Acessível de `http://www.portalingua.info/fr/poids-des-langues/`

[8] Por exemplo, os indicadores relativos à Internet têm origem numa companhia, provavelmente americana, que à data da escrita (e consequente consulta do seu sítio, no dia 2 de janeiro de 2012), tem em lugar proeminente o anúncio "Date sexy African women", o que não abona em favor da seriedade da mesma – embora não sejam necessariamente responsáveis pelo conteúdo da publicidade, o facto de terem publicidade, na minha opinião, reduz a impressão geral de confiabilidade.

[9] `http://www.observatorio-lp.sapo.pt/`

[10] `http://www.unesco.org/xtrans/bscontrib.aspx`

[11] Os valores referidos foram lidos em 3 de janeiro de 2012 dos seguinte endereço: `http://dtil.unilat.org/LI/2007/pt/cultura_letras_pt.htm`.

e encorajar os leitores a tomá-lo em conta.

Já em Aires e Santos (2002) tentámos estimar o tamanho da rede em português, mas nessa altura não estávamos interessadas em comparações com outras línguas, e por isso os nossos resultados não podem ser invocados como dados relativos.

De qualquer maneira, não só os estudos lusófonos não são muito difundidos, como as principais fontes quantitativas sobre os mesmos não nos inspiram suficiente confiança, o que demonstra a urgência de neles insistirmos e a eles dedicarmos os nossos esforços.

5 Estudos de utilizadores

É bem conhecida a dificuldade de avaliar, em termos de utilizadores, as vantagens e desvantagens de um sistema, sobretudo quando se trata de um sistema interativo.

O nosso caso, de tentar avaliar um recurso com o qual se interage, ainda é mais complexo, por duas razões:

- Para garantir a comparabilidade, desenvolvemos um método diferente (e pior) de interagir com a Wikipédia, que não só impedia o utilizador de ter acesso a imagens e a tabelas, como limitava a sua navegação para escolher e marcar justificações, o que torna o estudo da interação fundamentalmente diferente da verdadeira interação com a wikipédia.
- Por outro lado, as condições de participação também não eram naturais: é pouco provável, pelo menos, encontrar utilizadores realmente interessados em todos os tópicos que oferecemos, o que é diferente de estudos feitos com utilizadores naturais e interessados numa tarefa comum.

Assim sendo, a tarefa do Págico torna-se artificial por duas razões...

De qualquer maneira, o trabalho de Su (Su, 1998) mostrou que uma avaliação que entre em conta com a impressão geral de uma tarefa (e de um conjunto de resultados) é mais indicada do que outra que apenas classifique e avalie um a um. Por isso, o facto de procurarmos informação sobre um tópico e não sobre respostas individuais pode ser meritório, sobretudo se conseguirmos quantificar medidas com base no todo e não apenas em resultados individuais.

Não podemos contudo deixar de afirmar que o Págico é uma gota de água no oceano, e que veio demonstrar, se precisássemos de demonstração, que a usabilidade e o desenho de sistemas que de facto ajudam as pessoas em tarefas reais é algo que requer uma atenção específica e um trabalho aturado, e que estamos muito longe de ter conseguido definir um ambiente em que, fora de uma competição específica e experimental, pudesse atrair de facto pessoas para a utilizar.

Contudo, uma das motivações do Págico – sobre a qual falaremos mais no artigo que faz o seu balanço, Santos et al. (2012) – era sair para a realidade dos falantes de português e não nos confinarmos sempre e apenas à comunidade estreita de desenvolvedores de sistemas de PLN ou RI.

Como motivação, é importante reafirmá-la e refletir em como a implementar, quiçá, de maneira diferente.

6 Construção de recursos e de competência

Ao embarcarmos no Págico, alimentava-nos pelo menos uma certeza: o trabalho que realizaríamos iria ser benéfico à comunidade porque criaríamos recursos públicos e fomentaríamos a discussão e a reflexão numa área ou conjunto de áreas que nos pareciam – e parecem – importantes.

Esse é um dos traços distintivos da Linguateca (Santos, 2009), e que aproveito para repetir aqui: a construção de recursos públicos e a dinamização de áreas de investigação e de desenvolvimento no processamento e uso da língua portuguesa.

7 Comentários finais

O Págico distinguiu-se das avaliações conjuntas anteriores da Linguateca principalmente por duas características: quase não teve colaboração dos participantes, ou seja, pese embora o adjetivo "conjunto", foi totalmente decidido pela equipa, com base no raciocínio descrito no presente artigo, e mais globalmente, na presente edição da Linguamática; e dedicou-se a uma tarefa que tentámos que fosse muito mais próxima do dia-a-dia de uma população falante do português, em vez de uma tarefa técnica, automática, de apoio. Com esta decisão tentámos afastar-nos do efeito protótipo e também abordar os problemas da interação pessoa-máquina.

Continuamos a achar que um sistema que permitisse encontrar uma série de respostas a uma dada necessidade de informação na wikipédia podia servir como um auxiliar poderoso em várias profissões relacionadas com o conhecimento, tal como jornalista, escritor de temas de divulgação,

e mesmo estudante de uma dada área ou assunto. Parece-nos também que um tal sistema podia ser uma peça fulcral numa tarefa maior, como a que outros investigadores se têm atrevido, nomeadamente a criação automática de novas páginas da wikipédia (Sauper e Barzilay, 2009; Balasubramanian e Cucerzan, 2009), ou de outras formas de visualização da mesma.

Por outro lado, esperamos que, como efeitos laterais desta iniciativa, possamos obter mais conhecimento sobre o problema e sobre a cultura em português que possam servir para aumentar os recursos que a Linguateca põe à disposição da comunidade em geral, tanto a do processamento computacional da língua portuguesa, como a mais geral dos interessados na cultura lusófona.

Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN, e em 2011, pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN).

O Págico, e em particular o meu trabalho nesta avaliação conjunta, foi também financiado pela Universidade de Oslo.

Agradeço aos organizadores do PROPOR o convite feito para integrar o Págico neste contexto, ao Fernando Perdigão especialmente, pelo empenho e entusiasmo em relação ao mesmo, e a toda a organização do Págico. Estou também especialmente grata à Belinda Maia e à Cláudia Freitas pelos comentários pertinentes sobre versões preliminares deste artigo.

Referências

Aires, Rachel e Diana Santos. 2002. Measuring the web in portuguese. Em Brian Matthews, Bob Hopgood, e Michael Wilson, editores, *Euroweb 2002 conference*, pp. 198–9, 17-18 de Dezembro, 2002.

Balasubramanian, Niranjan e Silviu Cucerzan. 2009. Automatic generation of topic pages using query-based aspect models. Em *CIKM'09*, pp. 2049–52, 2-6 de Novembro, 2009.

Bucher, Bénédicte, Paul Clough, Hideo Joho, Ross Purves, e A. K. Syed. 2005. Geographic IR Systems: Requirements and Evaluation. Em *Proceedings of the 22nd International Cartographic Conference ICC 2005*, 11-16 de Julho, 2005.

Calvet, Louis-Jean. 2006. *Towards an ecology of world languages*. Polity, Londres.

Calvet, Louis-Jean. 2008. Le 'poids' des langues: une présentation de la situation linguistique du monde à l'heure de la mondialisation. Colloque de l'ACFAS 20008, Assocation des universités de la francophonie canadienne, `http://www.francophoniecanadienne.ca/DATA/ANNONCE/159.pdf`.

Dorow, Beate. 2006. *A Graph Model for Words and their Meanings*. Tese de doutoramento, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart.

Freitas, Cláudia, Paulo Rocha, Cristina Mota, Luís Costa, e Diana Santos. 2012. O que é uma resposta? Notas de uns avaliadores estafados. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Giampiccolo, Danilo, Pamela Forner, Anselmo Peñas, Christelle Ayache, Dan Cristea, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu, e Richard Sutcliffe. 2008. Overview of the CLEF 2007 Multilingual Question Answering Track. Em Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras, e Diana Santos, editores, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152, pp. 200–236, Berlim, Setembro, 2008. Springer.

Hagen, Arnulf. 2008. Battlefield wikipedia. Em Dagbladet.no. 21 de Outubro, `http://www.dagbladet.no/2011/10/21/kultur/debatt/kronikk/wikipedia/manipulasjon/18716526/`.

Heyer, Gerhard, Florian Holz, e Sven Teresniak. 2009. Change of topics over time and tracking topics by their change of meaning. Em Ana L. N. Fred, editor, *KDIR 2009: Proc. of Int. Conf. on Knowledge Discovery and Information Retrieval*. INSTICC Press, Outubro, 2009.

Heyer, Gerhard, Daniel Keim, Sven Teresniak, e Daniela Oelke. 2011. Interaktive explorative Suche in großen Dokumentbeständen. *Datenbank-Spektrum*, 11(3):195–206, Outubro, 2011. 10.1007/s13222-011-0072-4.

Jijkoun, Valentin e Maarten de Rijke. 2007. WiQA: Evaluating Multi-lingual Focused Access to Wikipedia. Em *The First International Workshop on Evaluating Information Access (EVIA)*, 15 de Maio, 2007.

Karlgren, Jussi, Paul Clough, e Julio Gonzalo. 2006. Multilingual Interactive Experiments with Flickr. *ERCIM News*.

Mota, Cristina. 2012. Resultados págicos: participação, medidas e pontuação. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Neto, Lauro. 2012. Wikipédia pede ajuda aos universitários. Em O Globo Educação, 4 de março, 2012. `http://oglobo.globo.com/educacao/wikipedia-pede-ajuda-aos-universitarios-4201782`.

Prado, Daniel, Daniel Pimienta, e Álvaro Blanco. 2009. *Douze années de mesure de la diversité linguistique sur Internet : bilan et perspectives*. Unesco.

2011. Págico: português mágico. Folheto de divulgação do Págico, `http://www.linguateca.pt/Pagico/Pagico.pdf`.

Ringel, Martin. 2011. IBM Watson and Jeopardy! Apresentação na conferência NoTur 2011, Oslo. `http://www.notur.no/notur2011/material/Martin_Watson_Notur_handout.pdf`.

Santos, Diana. 2007. Avaliação conjunta. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, capítulo 1, pp. 1–12.

Santos, Diana. 2009. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, 1(1):25–59, Maio, 2009.

Santos, Diana e Luís Miguel Cabral. 2009. Summing GikiCLEF up: expectations and lessons learned. Em Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, Giovanna Roda, Francesca Borri, Alessandro Nardi, e Carol Peters, editores, *Multilingual Information Access Evaluation, Vol. I: Text Retrieval Experiments*, volume Vol. I: Text Retrieval Experiments, pp. 212–222, Berlim / Heidelberg. Springer.

Santos, Diana, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Julia Schulz, Yvonne Skalban, e Erik Tjong Kim Sang. 2010. GikiCLEF: Crosscultural Issues in Multilingual Information Access. Em Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, Maio, 2010. European Language Resources Association (ELRA).

Santos, Diana, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, e Yvonne Skalban. 2009. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer, pp. 894–905.

Santos, Diana, Cristina Mota, Alberto Simões, Luís Costa, e Cláudia Freitas. 2012. Balanço do Págico e perspetivas de futuro. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Saracevic, T. 1995. Evaluation of evaluation in information retrieval. Em *Proceedings of the 17th Annual International ACM SIGIR'95, Conference on Research and Development in Information Retrieval*, pp. 138–146, Seattle, WA, EUA.

Sauper, Christina e Regina Barzilay. 2009. Automatically Generating Wikipedia Articles: A Structure-Aware Approach. Em *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp. 208–216, 2-7 de Agosto, 2009.

Speer, Robert, Catherine Havasi, e Henry Lieberman. 2008. AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. Em *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008*, pp. 548–553, 13-17 de Julho, 2008.

Su, Louise T. 1998. Value of search results as a whole as the best measure of information retrieval performance. *Information Processing and Management*, 34:557–579.

Thompson, Clive. 2010. Smarter Than You Think: What Is I.B.M.'s Watson? Em

New York Times, 20 de Junho, 2010. http://www.nytimes.com/2010/06/20/magazine/20Computer-t.html?_r=1&ref=homepage&src=me&pagewanted=all.

União Latina – Direção de Terminologia e Indústrias da Língua (DTIL). s.d. Línguas e culturas na web: Estudo 2007. Relatório técnico, União Latina. http://dtil.unilat.org/LI/2007/index_pt.htm.

Veale, Tony. 2007. Enriched Lexical Ontologies: Adding new knowledge and new scope to old linguistic resources. Curso na ESSLLI 2007, Dublin, Irlanda.

Widdows, Dominic. 2004. *Geometry and meaning*. CSLI Publications.

A lusofonia na Wikipédia em 150 tópicos

Cláudia Freitas
Linguateca/FCCN & PUC-Rio
maclaudia.freitas@gmail.com

Resumo

Este artigo descreve os tópicos usados no Págico, a primeira avaliação conjunta em recolha de informação centrada em tópicos relacionados com a lusofonia, usando o material da Wikipédia em português. Depois de uma explicação sobre como os tópicos foram escolhidos e de questões associadas à sua escolha e à sua categorização posterior, os tópicos são apresentados por categoria. Comentamos também a forma de apresentação de certas informações na Wikipédia e, por fim, fazemos uma categorização geográfica e temporal dos tópicos.

Palavras chave

Págico, Wikipédia

1 Apresentação

Este artigo descreve os tópicos usados no Págico, a primeira avaliação conjunta em recolha de informação centrada em tópicos relacionados com a lusofonia, usando o material da Wikipédia em português.

Depois de uma explicação de como os tópicos foram escolhidos[1], e de várias questões associadas à sua escolha e categorização posterior, comentamos algumas opções, decorrentes da forma pela qual certas informações estão presentes na Wikipédia e das características da própria tarefa. Por fim, fazemos uma caracterização dos tópicos em termos categorização por tema, por distribuição geográfica e por tempo.

2 Processo de elaboração e escolha dos tópicos

Já estamos habituados a usar buscadores como o Google para encontrar a informação que desejamos, embora nem sempre a tarefa seja fácil ou bem-sucedida. E em que contextos visitamos a Wikipédia? Para qual tipo de busca?

De minha experiência pessoal, não vou à Wikipédia como quem vai ao Google, isto é, procurar coisas. Vejo a Wikipédia como um lugar para "aprofundar superficialmente" pontos que me interessam. De alguma maneira, já sei o que procuro, já sei por onde começar.

Por isso, a tarefa de elaborar perguntas / tópicos que envolvessem busca não-trivial na Wikipédia acabou sendo mais difícil do que o esperado. (Aliás, do que eu esperava). Que tipo de perguntas elaborar cujo processo de busca fosse "não-trivial"?

A ideia era formular questões suficientemente interessantes – para falantes lusófonos – e que, por outro lado, tivessem respostas pouco óbvias ou cuja busca fosse trabalhosa (isto é, aquelas que, preferencialmente, tivessem respostas "espalhadas" por diferentes páginas.) Assim, diferentemente do relatado em (Simões, Rocha e Fonseca, 2009) sobre os diferentes tipos de pergunta comuns em WebPapers (competição em que os participantes buscam, na Web, respostas a perguntas), evitamos perguntas com respostas únicas, bem como perguntas com muitas indireções, do tipo "Qual a capital da quarta província mais populosa do Canadá").

Uma pergunta/tópico como [Músicos associados ao surgimento da Bossa-nova], por exemplo, seria uma pergunta relativamente fácil, pois rapidamente respondida com uma visita à página "Bossa Nova" da Wikipédia (ainda que, do ponto de vista dos sistemas, encontrar a resposta correta esteja longe de ser uma tarefa simples).

A estratégia esteve, portanto, em privilegiar perguntas consideradas difíceis/trabalhosas do ponto de vista dos humanos. Mesmo que essa não tenha sido a melhor opção, é defensável quando se pensa em uma tarefa o menos artificial possível.

Ainda na perspectiva de naturalizar a tarefa, outro aspecto relevante durante a elaboração das perguntas/tópicos foi conciliar perguntas cuja resposta não fosse facilmente obtida por uma busca no Google, o que tornaria artificial a necessidade de procura na Wikipédia.

Assim, embora subjacente a essa estratégia

[1] O grupo responsável pela elaboração dos tópicos foi constituído por mim, Paulo Rocha e Diana Santos.

esteja a mistura de duas tarefas - resposta automática a perguntas/recuperação de informação com a tarefa mais simples de recuperação de documentos - o que eu esperava era garantir que a busca de respostas - pelos participantes humanos ou sistemas - seria em grande medida dependente do conteúdo da Wikipédia.

Tentei, assim, criar perguntas que evitassem (ou minimizassem) uma busca geral na Internet, para posterior mapeamento, na Wikipédia, da página referente à resposta.

Naturalmente essa estratégia pode ser vista como complementar: a busca geral na Internet fornece documentos com a indicação da resposta, e o conteúdo da Wikipédia confirma/aprofunda os resultados obtidos. Mas, para os objetivos da tarefa proposta no Págico, tentei criar perguntas cuja resposta não fosse tão óbvia para o Google.

Vale notar, entretanto, que nem todas as perguntas foram guiadas por essas estratégias - mas isso contou para cerca de 60% delas.

Além de tentar elaborar tópicos interessantes e com resposta trabalhosa, evitamos perguntas com uma única resposta, justamente para motivar a procura em diferentes páginas.[2] E, a fim de evitar respostas abrangentes/amplas demais, ou com informação subjetiva, fugimos de tópicos/perguntas que envolvessem julgamentos de valor, do tipo "os melhores...".

3 Dificuldades

Assumindo que o Págico só teria perguntas que pudessem ser respondidas por meio da consulta à Wikipedia, tentamos já fornecer respostas à medida que elaboramos as perguntas como garantia de que estávamos diante de uma pergunta válida. Nesse processo, fomos surpreendidos por páginas cuja ligação não apontava para o termo referido - ou por erro, ou por inexistência da página. Comento a seguir dois desses casos:

tópico 52 [Gêneros musicais que misturam samba e ritmos norte-americanos]. Em uma série de páginas, existe referência à Pilantragem (que seria uma resposta adequada):

> *... e que foi chamado de pilantragem (uma mistura de samba e soul), movimento também idealizado e capitaneado por Carlos Imperial.*
>
> Wikipédia

A ligação de pilantragem, no entanto, leva à página *Turma da Pilantragem* (Wikipédia), que não é uma resposta adequada à pergunta:

> *A Turma da Pilantragem foi o nome de um grupo musical surgido no movimento cultural brasileiro denominado Pilantragem, em fins da década de 1960.*
>
> Turma da Pilantragem
> Wikipédia

Ou seja, pilantragem, como gênero musical, não tem uma página específica, embora, na página do referido grupo musical existam dados históricos sobre o movimento.

tópico 62 [Praias de Portugal boas para a prática de surf]. Na página *Turismo em Portugal* (Wikipédia), temos que

> *Portugal é também um país onde se pratica, além de muitos outros desportos, surf. Entre os melhores spots estão o Guincho, Peniche, Ericeira, Carcavelos, São Pedro e São João do Estoril, Costa da Caparica e São Torpes.*
>
> Turismo em Portugal
> Wikipédia

No entanto, a página/documento *São Torpes* (Wikipédia) não existe, e a ligação de São Torpes é para a página do *Parque Natural do Sudoeste Alentejano e Costa Vicentina* (Wikipédia). Na página do referido parque, por sua vez, não há qualquer menção a São Torpes. Assim, embora seja uma resposta correta, "São Torpes" não pode ser adicionada como página-resposta do Págico (e nem aceitamos a página *Parque Natural do Sudoeste Alentejano e Costa Vicentina* (Wikipédia), pois não consideramos que seja uma resposta adequada para o tópico.)

Ainda no mesmo trecho, a ligação de *Guincho* (Wikipédia) nos direciona para a página *Forte do Guincho* (Wikipédia), que também não consideramos correta - embora exista a página *Praia do Guincho* (Wikipédia), que responde ao tópico. Para a discussão referente à adequação

[2] De acordo com a Diana,

> à medida que o tempo passava, os critérios de escolha e distribuição foram sendo mais pelo interesse pelo aumento do leque de assuntos e menos pela existência de respostas na wikipédia. Associado ao fato infeliz de todo o trabalho ter sido feito sob grande pressão de prazos, isso explica que em alguns casos os donos dos tópicos (eu, por exemplo) não foram capazes ou não tiveram tempo de achar respostas antes de o conjunto estar completo e divulgado aos participantes.

de uma resposta, bem como de justificativas associadas às respostas, veja-se (Freitas et al., 2012).

Ou seja, tivemos dois tipos de problemas:

1. a ligação aponta para uma página incorreta, e a página correta ainda não existe na Wikipédia
2. a ligação aponta para uma página incorreta, e a página correta existe na Wikipédia

3.1 Redação dos tópicos

Como já mencionado, evitamos tópicos sem resposta – ou melhor, sem página de resposta. Tentamos evitar, também, tópicos com um número excessivo de respostas. Tópicos como [Orquestras filarmônicas com mais de cinquenta anos de vida],[Violoncelistas de língua portuguesa] e [Excetuando-se o português, que outras línguas faladas são faladas em países lusófonos], por exemplo, passaram respetivamente a **tópico 3** [Orquestras filarmónicas com mais de cinquenta anos de vida em países lusófonos] **tópico 4** [Mulheres violoncelistas de língua portuguesa] e a **tópico 25** [Que línguas bantas ou bantus são faladas em países lusófonos?]

Outro cuidado esteve em procurar um "português universal", isto é, garantir que diferenças entre as variantes não levariam a dificuldades na compreensão dos tópicos. Em alguns casos, tratava-se apenas de uma variação no uso (futebolistas / jogadores de futebol), não sendo difícil perceber do que se tratava. Noutros casos, embora fosse possível inferir a intenção da pergunta, as alterações foram necessárias para um rápido entendimento comum. Por exemplo, [Ilhas e ilhéus desabitados de Cabo Verde] passou a "Ilhas e ilhotas", uma vez que, no português falado no Brasil, *ilhéu* é principalmente aquele que mora em uma ilha, o que tornava o tópico sem sentido.

4 Conteúdo e distribuição dos tópicos

Considerando o espetro assumidamente geral de "cultura lusófona", e também que um tópico pode ser sobre mais de um tema ([Filmes sobre as relações entre Portugal e suas colônias], por exemplo, pode ser tanto sobre Cinema quanto História), os 150 tópicos distribuem-se da maneira indicada na tabela 1.

Esta tabela apresenta a negrito a distribuição dos tópicos pelo que chamamos grandes temas – categorias mais abrangentes. A maior parte dos 150 tópicos pertence ao super tema Letras, decorrência principalmente da presença do tema História – que corresponde ao tipo de informação mais naturalmente associada ao conhecimento enciclopédico –nessa categoria. A segunda categoria mais frequente é Artes, que engloba tópicos relacionados a música e cinema, dentre outros. Logo em seguida aparecem o super tema Geografia e Cultura. Os super temas menos frequentes foram Política, Desporto/Esportes, Ciência e Economia.

Grande tema	Tópicos	
-tema	#	%
Letras	**69**	**46,00**
- história	50	33,33
- literatura	15	10,00
- linguística	6	4,00
- jornalismo	3	2,00
- filosofia	2	1,33
Artes	**36**	**24,00**
- música	19	12,67
- cinema	10	6,67
- televisão	4	2,67
- artes plásticas	2	1,33
- artes	2	1,33
Geografia	**34**	**22,67**
- geografia	26	17,33
- arquitetura/urbanismo	7	4,67
- demografia	4	2,67
- geologia	2	1,33
Cultura	**27**	**18,00**
- antropologia/folclore	12	8,00
- religião	7	4,67
- culinária	5	3,33
- cultura	3	2,00
- ensino	2	1,33
Política	**19**	**12,67**
Desporto/Esportes	**18**	**12,00**
Ciência	**14**	**9,33**
- saúde	4	2,67
- zoologia	3	2,00
- ciência	2	1,33
- botânica	2	1,33
- geologia	2	1,33
- matemática	1	0,67
Economia	**6**	**4,00**

Tabela 1: Distribuição dos tópicos por tema

Como se pode perceber, a totalidade de tópicos distribuídos é superior a 150, indicando que alguns tópicos foram classificados em mais de um grande tema. Mais especificamente, à maioria dos tópicos (56%) foi atribuído apenas um tema; a 39% dos tópicos foram atribuídos 2 temas e a 4,6% dos tópicos, 3 temas; não há tópicos com mais de 3 temas (veja-se a tabela 2).

# Temas atribuídos	Grandes temas		Temas	
	Total	%	Total	%
1	84	56,00	75	50,00
2	59	39,33	64	42,67
3	7	4,67	11	7,33

Tabela 2: Atribuição de temas e grandes temas por tópico

A tabela 1 mostra igualmente, abaixo da distribuição por grandes temas, a distribuição dos tópicos por temas, o que permite uma análise mais detalhada do conteúdo do Págico. A partir desses valores observamos que, dentre os temas, o mais frequente foi História (50 tópicos), seguido de Geografia (26 tópicos), Música e Política (ambos com 19 tópicos) e Desporto/Esportes (18 tópicos). O menos frequente foi Matemática, com apenas um tópico. A tabela 2 indica também que metade dos tópicos é apenas sobre um tema; 42% (64 tópicos) dos tópicos envolve 2 temas e apenas 7% (11) dos tópicos foi classificado como 3 temas. Contrastando a distribuição por tópicos em termos de temas e super temas (consulte-se novamente a tabela 2), notamos que, quando tratamos das categorias mais abrangentes (super temas), há uma ligeira prevalência de tópicos com um tema (56% vs 50%). A atribuição de dois ou mais temas a um tópico foi decorrência não só do conteúdo misto do tópico propriamente, como no exemplo de filmes sobre um dado tema, mas também quando houve divergência, entre os membros da organização , sobre qual tema atribuir – e, nesse caso, todas as possibilidades foram consideradas. Por essa perspectiva, a maior frequência de tópicos com categorização única no âmbito dos super temas sugere a maior facilidade de concordância quando estamos diante de categorias mais amplas.

A tabela 5 lista todos os tópicos criados para o Págico. Para cada tópico, informamos: identificador (ID), descrição, a sua classificação em grande tema, tema e local (consulte-se a tabela 4 para uma correspondência entre as abreviaturas e os países respetivos).

4.1 Entidades Mencionadas nos tópicos

Dos 150 tópicos, 96 continham entidades mencionadas em sua formulação. A tabela 3 apresenta a distribuição de entidades mencionadas (EM) por tópico. Para a identificação e classificação das EM, seguimos as diretrizes do Segundo HAREM (Mota e Santos, 2008): consideramos apenas as EM iniciadas por letras maiúsculas – exceção para as entidades temporais – e tomamos como base as categorias descritas em Carvalho et al. (2008), uma vez que se mostraram produtivas para a classificação das entidades nesta tarefa. Ainda seguindo a maneira de classificação do Segundo HAREM, consideramos possível uma mesma EM receber mais de uma classificação. Assim, a EM Brasil no **tópico 34** [Viajantes ou exploradores que escreveram sobre o Brasil do século XVI] pode ser vista tanto como uma ORGANIZAÇÃO, como um LOCAL ou como PESSOA (os brasileiros), e portanto recebeu as 3 classificações.

local	59
tempo	24
organização	21
pessoa	10
acontecimento	8
abstração	5
coisa	3

Tabela 3: EM no Págico

4.2 Categorização geográfica e temporal

Com relação à localização geográfica, os 150 tópicos do Págico estão distribuídos como indica a tabela 4:

Lugar (abreviatura)	Tópicos	
	#	%
Brasil (br)	50	33,33
Lusofonia (lus)	44	29,33
Portugal (pt)	16	10,67
Moçambique (mo)	11	7,33
Angola (ao)	8	5,33
Geral (ger)	7	4,67
Cabo Verde (cv)	6	4,00
Macau (mo)	4	2,67
Timor (ti)	3	2,00
Guiné Bissau (gw)	2	1,33
São Tomé e Príncipe (st)	2	1,33

Tabela 4: Lugares mencionados no Págico

A categoria "lusofonia geral" corresponde a tópicos em que não há especificação de região geográfica; e os tópicos "gerais" não mencionam explicitamente elementos da cultura lusófona, mas integram o Págico por conterem, em sua resposta, elementos da cultura de países lusófonos.

Com relação à localização temporal, os 150 tópicos estão assim distribuídos: a partir do século XX em diante, existem 43, até o século XIX inclusive, 13, enquanto os outros não envolvem localização temporal explícita, quer por corresponderem a perguntas que não envolvem tempo (1), ou perguntas cujas respostas podem estar distribuídas por diferentes períodos (2):

1. **tópico 123** [Rios de Angola com mais de 500 quilómetros de comprimento]
2. **tópico 103** [Movimentos culturais surgidos no nordeste do Brasil]

4.3 E os tópicos de exemplo?

Rodrigues, Gonçalo Oliveira e Gomes (2012) mencionam que os tópicos do Págico eram muito diferentes dos tópicos de exemplo. Não tínhamos essa impressão, mas fomos aplicar a análise acima a esses 11 tópicos:

Comparativamente aos 11 tópicos apresentados como exemplo, os tópicos do Págico têm uma distribuição bastante parecida em temos de localização geográfica. No critério conteúdo, como era de se esperar, há uma distribuição mais homogênea dos temas por tópico. Os supertemas mais frequentes foram Artes, Ciência, Cultura, Desporto/Esportes, todos com 2 tópicos cada, e Economia, Geografia, Letras e Política, com 1 tópico cada. Quanto à distribuição geográfica, dos 11 tópicos, 5 (45%) estão relacionados ao Brasil; 3 relacionados a Portugal ; 2 à Lusofonia geral e 1 é um "tópico geral". Dos 11 tópicos, apenas 6 (cerca de 55%) contêm entidades mencionadas, em uma distribuição parecida à do Págico (64% dos tópicos contêm EM). Nos exemplos, a distribuição das categorias de EM se deu da seguinte maneira (lembrando sempre que uma EM pode receber mais de uma categoria): ORG e LOCAL foram as mais frequentes, com 2 ocorrências cada, e em seguida PESSOA, TEMPO e OUTRO, com 1 ocorrência cada. Por fim, considerando a dimensão temporal, assim como no Págico, os tópicos exemplo tratam, na imensa maioria (7 tópicos), de questões relativas ao século XX.

5 Considerações finais

Neste breve artigo apresentamos as motivações e decisões subjacentes à criação dos tópicos do Págico. Caracterizamos detalhadamente o conjunto de tópicos quanto ao tema, distribuição geográfica e temporal, em um instantâneo que tenta capturar, em 150 perguntas, um pouco da cultura dos países falantes da língua portuguesa.

Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN, e em 2011 pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN).

Agradeço ao resto da organização do Págico, sem a qual o mesmo não teria sido possível, e a todos os participantes, cujas respostas ajudaram a iluminar os tópicos e a esclarecer pontos pouco claros.

Agradeço também ao Fernando Perdigão e à Belinda Maia, pela revisão e comentários que ajudaram a tornar o artigo mais esclarecedor.

Referências

Carvalho, Paula, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas, e Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 11–31, 31 de Dezembro, 2008.

Freitas, Cláudia, Paulo Rocha, Cristina Mota, Luís Costa, e Diana Santos. 2012. O que é uma resposta? Notas de uns avaliadores estafados. *Linguamática*, 4(1), Abril, 2012. Neste volume.

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

Rodrigues, Ricardo, Hugo Gonçalo Oliveira, e Paulo Gomes. 2012. Uma abordagem ao Págico baseada no processamento e análise de sintagmas dos tópicos. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Simões, Alberto, Paulo Rocha, e Rúben Fonseca. 2009. Webpaper — más perguntas e boas respostas: a arte de interrogar. Em Paulo Dias, António José Osório, e Altina Ramos, editores, *O digital e o currículo*. Centro de Competência da Universidade do Minho, pp. 227–238, Maio, 2009.

Tabela 5: Lista de tópicos do Págico.

ID	Descrição	Grande tema	Tema	Lugar
1	Filmes sobre a ditadura ou sobre o golpe militar no Brasil	Artes, Letras	cinema, história	br
2	Telenovelas brasileiras do gênero realismo fantástico.	Artes	televisão	br
3	Orquestras filarmónicas com mais de cinquenta anos de vida em países lusófonos	Artes	música	lus
4	Mulheres violoncelistas de língua portuguesa	Artes	música	lus
5	Flautistas que se naturalizaram brasileiros ou portugueses.	Artes	música	lus
6	Instrumentistas famosos, de países de língua portuguesa, que são conhecidos por tocarem bem mais do que um instrumento.	Artes	música	lus
7	Guitarristas portugueses que também foram compositores.	Artes	música	pt
8	Telenovelas brasileiras passadas no tempo da escravatura no Brasil	Artes, Letras	televisão, história	br
9	Comidas de santo (comidas rituais do Candomblé ou Umbanda) que também fazem parte da culinária brasileira.	Cultura	culinária, religião, antro./fol.	br
10	Pratos brasileiros de origem ou influência indígena.	Cultura, Letras	culinária, história	br
11	Filmes sobre o cangaço.	Artes, Letras	cinema, história	br
12	Políticos de movimentos de libertação lusófonos que foram recebidos pelo Papa	Letras, Política	política, história	lus
13	Dinossauros carnívoros que habitaram o Brasil.	Ciência	zoologia	br
14	Filmes baseados em histórias de Guimarães Rosa.	Artes, Letras	cinema, literatura	br
15	Filmes brasileiros sobre futebol.	Artes, Esportes	cinema, esportes	br
16	Membros da igreja associados à Teologia da Libertação.	Cultura, Letras, Política	filosofia, religião, política	ger
17	Documentários sobre políticos brasileiros.	Artes, Letras, Política	cinema, política, história	br
18	Discos brasileiros considerados marcantes na história da música brasileira	Artes	música	br
19	Tribos indígenas que vivem na Amazônia.	Geografia	demografia, geografia	br
20	Que etnias africanas entraram no Brasil no período da escravatura?	Geografia, Letras	história, demografia	br
21	Lendas e figuras folclóricas de origem portuguesa populares no Brasil	Cultura, Letras	antro./fol., história	lus
22	Filmes sobre o Brasil colonial	Artes, Letras	cinema, história	br
23	Filmes sobre as relações entre Portugal e suas colônias	Artes, Letras	cinema, história	lus
24	Filmes que abordem movimentos históricos como rebeliões, revoltas, batalhas, levantes ou levantamentos populares entre a população (ou parte dela) e o governo vigente em países lusófonos	Artes, Letras, Política	cinema, história, política	lus
25	Que línguas bantas ou bantus são faladas em países lusófonos?	Letras	linguística	lus
26	Artistas brasileiros cujas obras integram o acervo do MoMA ou que já expuseram lá.	Artes	artes plásticas	br
27	Doenças letais comuns em países lusófonos transmitidas por mosquitos	Ciência	saúde	lus
28	Festas populares brasileiras de origem portuguesa	Cultura	antro./fol.	lus
29	Escritores lusófonos que se filiaram a partidos políticos	Letras, Política	literatura, política	lus

Continua na página seguinte...

Tabela 5 – continuação da página anterior

ID	Descrição	Grande tema	Tema	Lugar
30	Batalhas entre brancos e índios brasileiros que ocorreram até o século XVIII	Letras	história	br
31	Grupos indígenas que habitavam o litoral do Brasil quando chegaram os europeus.	Geografia, Letras	história, demografia, geografia	br
32	Tribos indígenas brasileiras que praticavam canibalismo em rituais religiosos.	Cultura, Letras	história, religião	br
34	Viajantes ou exploradores que escreveram sobre o Brasil do século XVI	Geografia, Letras	história, geografia, literatura	br
35	Que autores não lusófonos escreveram sobre o Brasil nos séculos XVIII e XIX?	Letras	história, literatura	br
36	Escolas de samba fundadas ou sediadas em morros cariocas.	Artes, Cultura, Letras	música, história, antro./fol.	br
37	Ritmos lusófonos que utilizam zabumba, triângulo e sanfona.	Artes	música	lus
38	Esportes integrados às Olimpíadas nos últimos 30 anos	Esportes	esportes	ger
39	Modalidades esportivas em que países lusófonos já ganharam medalha de ouro nos Jogos Olímpicos.	Esportes	esportes	lus
40	Instrumentos musicais de origem africana comuns no Brasil	Artes, Letras	música, história	br
41	Congressos ou conferências que têm por tema as relações culturais e/ou sociais entre África e demais países lusófonos	Ciência, Cultura, Política	ciência, política, cultura	lus
42	Plantas com as quais índios brasileiros pintam seus corpos	Ciência, Cultura	botânica, antro./fol.	br
43	Produtos agrícolas com os quais se pode produzir combustível em escala comercial	Ciência, Economia	ciência, economia	ger
44	Lendas ou personagens folclóricas de origem indígena conhecidas no Brasil	Cultura	antro./fol.	br
45	Cantores vaiados nos festivais de música brasileira na década de 60.	Artes, Letras	música, história	br
46	Séries ou minisséries de televisão passadas na época da independência do Brasil	Artes, Letras	televisão, história	br
47	Filósofos franceses do século XX que deram cursos ou conferências no Brasil.	Letras	filosofia	br
48	Jornais portugueses que existiam no tempo da implantação da república	Letras	história, jornalismo	pt
49	Séries ou minisséries brasileiras baseadas em romances portugueses.	Artes, Letras	televisão, literatura	lus
50	Jornais que circularam no Rio de Janeiro entre 1910 e 1960.	Letras	história, jornalismo	br
51	Além do samba, que outros gêneros musicais são populares no carnaval brasileiro?	Artes, Cultura	música, antro./fol.	br
52	Gêneros musicais que misturam samba e ritmos norte-americanos.	Artes	música	ger
53	Parques do Rio de Janeiro que têm cachoeiras	Geografia	geografia	br
54	Igrejas do Rio de Janeiro construídas por irmandades ou confrarias de negros.	Geografia, Letras	arq./urb., história	br
55	Escritores estrangeiros que visitaram Portugal no século XIX e que publicaram descrições das suas viagens	Letras	literatura, história	pt
56	Jogadores de futebol brasileiros que foram jogar no exterior quando tinham menos de 21 anos de idade.	Esportes	esportes	br
57	Jogadores lusófonos que já foram campeões mundiais por clubes europeus.	Esportes	esportes	lus
58	Países que venceram a Copa do Mundo em uma disputa de pênaltis	Esportes	esportes	ger
59	Jogadores de basquetebol brasileiros que jogam ou jogaram em campeonatos da NBA	Esportes	esportes	br

Continua na página seguinte...

Tabela 5 – continuação da página anterior

ID	Descrição	Grande tema	Tema	Lugar
60	Jogadores de basquetebol brasileiros com mais de 2, 10 metros	Esportes	esportes	br
61	Movimentos culturais em países lusófonos que se refletiram nas artes plásticas e na música	Artes	artes plásticas, música	lus
62	Praias de Portugal boas para a prática de surf	Esportes, Geografia	esportes, geografia	pt
63	Estudiosos da música indígena brasileira	Artes, Cultura	música, antro./fol.	br
64	Doces brasileiros que têm origem nos doces portugueses	Cultura, Letras	culinária, história	lus
66	Fotógrafos de cinema ou diretores de fotografia brasileiros que já dirigiram ou realizaram filmes	Artes	cinema	br
67	Doenças que acometeram a população indígena nas Américas	Ciência, Letras	história, saúde	ger
68	Bandas brasileiras de punk formadas até 1980 em São Paulo.	Artes, Letras	música, história	br
69	Complexos esportivos construídos para os Jogos Pan Americanos do Rio de Janeiro (2007) que nunca receberam quaisquer outros eventos esportivos.	Esportes, Geografia	esportes, arq./urb.	br
70	Centros culturais e faculdades do Rio de Janeiro sediados em prédios históricos	Geografia	arq./urb.	br
71	Doenças presentes no Brasil no século XVII	Ciência, Letras	história, saúde	br
72	Autores lusófonos que escrevem literatura fantástica e que tenham pelo menos um livro publicado	Letras	literatura	lus
74	Nomes ligados à luta contra o racismo no século XX no Brasil	Letras, Política	política, história	br
75	Organizações ou grupos armados que lutaram contra o regime militar no Brasil	Letras, Política	política, história	br
76	Ritmos brasileiros de origem portuguesa	Artes, Letras	música, história	lus
77	Médicos portugueses do século XVIII que viveram a maior parte da sua vida no estrangeiro	Ciência, Letras	história, saúde	pt
78	Escritoras de língua portuguesa que tenham publicado livros para crianças entre 1850 e 1940	Letras	literatura, história	lus
79	Povos indígenas brasileiros considerados extintos.	Geografia, Letras	história, demografia	br
80	Línguas faladas em Timor Leste	Letras	linguística	ti
81	Igrejas em Macau	Cultura, Geografia	religião, arq./urb.	mo
82	Políticos timorenses que participaram na luta armada contra a Indonésia	Letras, Política	política, história	ti
83	Que equipes da primeira divisão do futebol brasileiro desceram para a segunda divisão e nunca mais conseguiram voltar?	Esportes	esportes	br
84	Escritores lusófonos traduzidos para 5 ou mais idiomas	Letras	literatura	lus
85	Destinos turísticos do Brasil cuja temperatura no Inverno pode ser negativa	Geografia	geografia	br
86	Compositoras brasileiras de samba	Artes	música	br
87	Sambistas negros que abordam o racismo em suas letras	Artes, Letras, Política	música, literatura, política	ger
88	Cidades portuguesas que têm festivais medievais	Cultura, Geografia	geografia, antro./fol.	pt
89	Excetuando o português, para que outras línguas consideradas oficiais em países lusófonos existe uma versão da Wikipédia?	Letras	linguística	lus
90	Filmes brasileiros premiados na categoria Montagem.	Artes	cinema	br
91	Estados fronteiriços de Moçambique	Geografia	geografia	mz

Continua na página seguinte...

Tabela 5 – continuação da página anterior

ID	Descrição	Grande tema	Tema	Lugar
92	Cidades que fizeram parte do domínio português na Índia	Geografia, Letras	história, geografia	pt
93	Estilos musicais cabo-verdianos	Artes	música	cv
94	Parques nacionais de Moçambique	Geografia	geografia	mz
95	Partidos políticos de São Tomé e Príncipe	Política	política	st
96	Montanhas mais altas de cada país lusófono	Geografia	geografia	lus
97	Escritores cabo-verdianos com obra publicada em crioulo	Letras	literatura, linguística	cv
98	Cidades dos Estados Unidos que tiveram forte imigração portuguesa	Geografia, Letras	história, geografia	pt
99	Matemáticos de língua portuguesa que estudaram ou trabalharam em Itália	Ciência	matemática	lus
100	Ilhas de Moçambique	Geografia	geografia	mz
101	Cidades e vilas em países não-lusófonos que se situem junto à fronteira de um país lusófono	Geografia	geografia	lus
102	Ordens religiosas que vieram para o Brasil na época colonial.	Cultura, Letras	história, religião	br
103	Movimentos culturais surgidos no nordeste do Brasil.	Cultura, Geografia	cultura, geografia	br
104	Pesquisadores do folclore brasileiro	Cultura	antro./fol.	br
105	Empresas de mineração angolanas	Economia	economia	ao
106	Vice-reis da Índia Portuguesa	Letras	história	pt
107	Dioceses católicas de Moçambique	Cultura	religião	mz
108	Jogadores de futebol nascidos em Cabo Verde que representaram a seleção portuguesa	Esportes	esportes	cv, pt
109	Candidatos a alguma das eleições presidenciais na Guiné-Bissau	Política	política	gw
110	Políticos da África lusófona que estudaram na União Soviética	Política	política	lus
111	Padres católicos que estão ou estiveram ativos em Timor	Cultura	religião	ti
112	Capitais das províncias de Angola	Geografia	geografia	ao
113	Ilhas e ilhotas de Cabo Verde que não são habitadas	Geografia	geografia	cv
114	Antigas moedas dos países lusófonos e suas colónias	Economia, Letras	economia, história	lus
115	Clubes que venceram o Girabola	Esportes	esportes	pt
116	Escritores lusófonos que passaram temporadas na prisão	Letras	literatura	lus
117	Produtos utilizados pelos portugueses no comércio de escravos com a África	Economia, Letras	história, economia	pt
118	Escritores moçambicanos que receberam o Prémio Camões	Letras	literatura	mz
119	Pratos típicos da gastronomia de Cabo Verde	Cultura	culinária	cv
120	Cervejas consumidas em Angola	Cultura	culinária	ao
121	Frutos de Angola	Ciência, Geografia	geografia, botânica	ao
122	Políticos lusófonos do século XX assassinados	Letras, Política	política, história	lus
123	Rios de Angola com mais de 500 quilómetros de comprimento	Geografia	geografia	ao
124	Cabo-verdianos que participaram na guerra colonial na Guiné	Letras, Política	história, política	gb, cv
125	Fortalezas e feitorias portuguesas na costa africana	Geografia, Letras	história, geografia, arq./urb.	pt
126	Atletas da África lusófona que tenham competido nos Jogos Olimpícos	Esportes	esportes	lus
127	Mamíferos herbívoros existentes em Moçambique	Ciência, Geografia	geografia, zoologia	mz
128	Escritores portugueses que tenham vivido em Macau	Letras	literatura	mo, pt
129	Antigos alunos da Universidade Eduardo Mondlane e da sua antecessora, a Universidade de Lourenço Marques	Cultura	ensino	mz

Continua na página seguinte...

Tabela 5 – continuação da página anterior

ID	Descrição	Grande tema	Tema	Lugar
130	Acordos, tratados e outros protocolos assinados entre as fações da Guerra Civil angolana	Letras, Política	história, política	ao
131	Quem descobriu São Tomé e Príncipe?	Letras	história	st
132	Deputados da FRELIMO	Política	política	mz
133	Futebolistas do Petro de Luanda	Esportes	esportes	ao
134	Personagens do século XIX ligadas à luta anti-colonial em Moçambique	Letras, Política	história, política	mz
135	Aves de Angola	Ciência, Geografia	geografia, zoologia	ao
136	Provas desportivas internacionais com participação da seleção moçambicana	Esportes	esportes	mz
137	Eventos onde Maria de Lurdes Mutola foi medalha de ouro	Esportes	esportes	mz
138	Jornais, revistas e outras publicações periódicas de Macau	Letras	jornalismo	mo
139	Infra-estruturas de transportes (aeroportos, estações, rodoviárias e ferroviárias, pontes, etc.) existentes em Macau	Geografia	arq./urb.	mo
140	Cidades lusófonas conhecidas pelo seu Carnaval	Cultura	antro./fol.	lus
141	Cidades das antigas colónias portuguesas que têm ou tiveram a designação de Nova	Geografia, Letras	geografia, linguística	lus
142	Diplomatas lusófonos que trabalharam no Japão	Letras, Política	história, política	lus
143	Cidades do império português ocupadas pelos holandeses no período filipino	Letras	história	pt
144	Locais referidos n Os Lusíadas	Letras	literatura	pt
145	Minas de países lusófonos em atividade durante mais de 30 anos consecutivos, ainda em funcionamento ou não	Ciência, Economia, Geografia	geografia, geologia, economia	lus
146	Vulcões em território português e brasileiro	Ciência, Geografia	geografia, geologia	lus
147	Museus em capitais de países lusófonos	Artes, Cultura	cultura, artes	lus
148	Primeiras universidades de cada país lusófono	Cultura, Letras	história, ensino	lus
149	Arquitetos de países lusófonos com obras em países estrangeiros na América do Norte e na Europa	Geografia	arq./urb.	lus
150	Empresários lusófonos com uma fortuna considerável	Economia	economia	lus
151	Cidades em países não lusófonos com mais de 500 mil habitantes ou com mais de 10% da populacão falando português	Geografia, Letras	geografia, linguística	lus
152	Pintores estrangeiros com uma ligacão forte a Portugal ou ao Brasil	Artes	artes	lus
153	Toureiros a cavalo de países lusófonos com carreira internacional	Cultura, Esportes	antro./fol., esportes	lus

Tirando o chapéu à Wikipédia: A coleção do Págico e o Cartola

Alberto Simões
Instituto de Letras e Ciências Humanas
Universidade do Minho
ambs@ilch.uminho.pt

Luís Costa
Linguateca/FCCN
luis.f.kosta@gmail.com

Cristina Mota
Linguateca/FCCN
cmota@ist.utl.pt

Resumo

Este artigo apresenta a coleção do Págico, ou seja, a coleção subjacente ao Págico, e o pacote de recursos do Págico, o Cartola, que inclui a própria coleção.

O artigo dá particular destaque à construção da coleção do Págico, uma coleção de documentos da Wikipédia portuguesa. Esta coleção foi criada com o objetivo de garantir (i) igualdade no recurso usado por todos os participantes, (ii) homogeneidade nas respostas e (iii) semi-automatização na avaliação das respostas. Em primeiro lugar, será justificada a necessidade da criação deste recurso. Posteriormente, serão apresentadas as alternativas existentes para a sua criação, qual a escolhida, e quais os problemas encontrados.

Além disso, o artigo caracteriza, segundo várias vertentes, a coleção do Págico bem como uma subcoleção desta, correspondente ao monte do Págico. O monte do Págico, também incluído no Cartola, inclui todas as respostas e justificações distintas encontradas pelos criadores de tópicos e pelos participantes.

Palavras chave

Wikipédia, Págico, Coleção, XHTML, Wiki

1 Introdução

Uma das grandes vantagens de uma avaliação conjunta é produzir um conjunto de recursos que podem ser usados no futuro para avaliar outros sistemas, estabelecendo uma bitola e a respetiva bancada de teste.

No decurso do Págico foi criado o Cartola, um pacote de recursos público constituído pela coleção do Págico (a coleção de documentos da Wikipédia de onde as respostas e justificações deviam ser escolhidas, primeiro, pelos criadores de tópicos e, depois, pelos participantes), por exemplos de tópicos e correspondentes respostas associadas às suas justificações, pelos tópicos de avaliação e pelas respostas dos criadores de tópicos e dos participantes com a respetiva avaliação feita pela organização. O Cartola é disponibilizado pela Linguateca em `http://www.linguateca.pt/Cartola/` e inclui especificamente:

- a coleção de documentos (689 629) da Wikipédia portuguesa usada no Págico;
- 11 exemplos de tópicos com as respetivas respostas e justificações (85);
- os 150 tópicos usados na avaliação;
- as corridas dos sistemas e as respostas dos participantes humanos em formato de corridas[1];
- o monte do Págico, ou seja, a coleção de todas as respostas com as suas justificações encontradas no Págico (quer pelos criadores de tópicos quer pelos participantes) e a respetiva avaliação.
- listas das respostas distintas corretas com (2 250) e sem as justificações (1 871);
- lista das respostas distintas corretas quer tenham sido bem ou mal justificadas, sem justificações (1 979);
- lista das respostas consideradas duvidosas.

Os tópicos de avaliação do Págico encontram-se descritos em (Freitas, 2012), enquanto (Freitas et al., 2012) discute a avaliação das respostas, analisando entre outras coisas as respostas duvidosas. Este artigo, por outro lado, foca a coleção do Págico e uma subcoleção desta, correspondente ao monte das respostas do Págico e que inclui, portanto, todos os documentos que foram usados como resposta ou justificação no Págico.

A coleção do Págico é uma coleção de documentos criada a partir de uma versão estática da Wikipédia portuguesa. Começaremos por justificar, na secção 2, a necessidade de criar este recurso. Discutiremos, em seguida, nas secções 3.1 e 3.2, um conjunto de definições e convenções usadas na Wikipédia que teria de ser estudado pelos participantes a fim de conseguirem processar de forma satisfatória as cópias disponibilizadas da

[1] Como referido em (Mota, 2012), a partir das respostas dadas no SIGA pelos participantes humanos foram criadas as corridas equivalentes.

Wikipédia. A secção 4 explica como o formato da Wikipédia foi processado e convertido num conjunto de documentos XHTML (que sendo um formato muito mais simples e amplamente usado facilita o processamento por parte dos participantes), o qual constitui a coleção do Págico.

Posteriormente, na secção 5, faremos uma caracterização da coleção do Págico e da subcoleção do monte do Págico de diferentes perspetivas. Esta caraterização irá permitir ao leitor ter uma noção da abrangência dos tópicos propostos para avaliação em relação à coleção como um todo. Além disso, permitirá que potenciais interessados no uso de coleções semelhantes em futuras avaliações fiquem com uma imagem do conteúdo real da Wikipédia portuguesa.

Terminaremos com algumas conclusões e propostas de melhoramentos para futuras edições, seja do Págico, seja de outras avaliações que usem a Wikipédia como fonte de informação.

2 Criar uma nova coleção, sim ou não?

A Wikipédia é um recurso em constante mutação. Por um lado, é o conteúdo que muda a cada instante, por outro, são as regras e a sintaxe que vão evoluindo. Esta constante mudança faz com que não seja um recurso fácil de usar para uma avaliação de qualquer tipo de ferramenta.

No caso concreto do Págico (consulte-se os restantes artigos nesta edição para mais informação sobre outros aspetos desta avaliação conjunta), em que se pretende avaliar ferramentas de recolha de informação na Wikipédia, este facto é de grande importância. No Págico, os participantes têm de encontrar artigos da Wikipédia que respondam a um tópico. Ora, se não existir uma versão estável, onde os participantes devam encontrar as ditas respostas, é possível que em determinado dia:

- exista um artigo que sirva de resposta (ou justificação) a um dos tópicos do Págico e que, no dia seguinte, esse artigo tenha desaparecido;
- não exista o artigo que sirva de resposta (ou justificação), mas no dia seguinte já tenha sido criado;
- exista um artigo que no dia seguinte é alterado de tal forma que invalida que seja uma resposta correta (ou que justifique adequadamente uma resposta).

Teria sido possível usar a coleção desenvolvida para o GikiCLEF (Santos et al., 2010), no entanto optou-se por usar uma versão mais recente da Wikipédia. Além do facto de garantir mais proximidade com a Wikipédia atual, também permite que possamos analisar (neste artigo) o estado da Wikipédia portuguesa. Infelizmente a abordagem usada para a construção da coleção para o GikiCLEF não foi possível de ser repetida já que a ferramenta usada já não é mantida.

Foi, então, necessário construir uma coleção estática que pudesse ser usada por todos os participantes, e que tornasse a avaliação mais simples (ou mesmo, possível). Para isso foi usada uma cópia estática da Wikipédia (a própria Fundação Wikimedia disponibiliza cópias regulares das várias versões da Wikipédia) de 25 de Abril de 2011[2].

Embora estas cópias estáticas da Wikipédia sejam disponibilizadas em vários formatos (como SQL, para introdução direta num gestor de bases de dados, ou num único documento XML com todos os artigos), esses formatos não são fáceis de processar, quer pelo seu tamanho, quer pelo próprio formato em que são disponibilizados, o que será discutido em seguida.

3 A Wikipédia

Todos conhecemos a Wikipédia, e já a consultámos pelo menos um par de vezes. No entanto, conhecemos a Wikipédia do ponto de vista de um utilizador comum, que consulta e lê um conjunto de artigos, e possivelmente não como um membro da comunidade da Wikipédia, tentando melhorar artigos, ou contribuir com novos artigos. Mesmo que já tenha editado um ou dois artigos da Wikipédia é natural que não tenha compreendido como a estrutura da Wikipédia é rica e, ao mesmo tempo, complexa.

A Wikipédia não é apenas um sistema *wiki* em que cada página corresponde a um artigo de uma enciclopédia. Existe uma estrutura de espaços de nomes (*namespaces*), entradas, entradas de desambiguação e de redireção, e macros e funções. Nesta secção apresentamos (de forma superficial) a estrutura e a sintaxe de macros e funções da Wikipédia relevantes à construção da coleção do Págico.

3.1 Estrutura da Wikipédia

A Wikipédia começou, como não podia deixar de ser, como um conjunto de páginas, em que cada uma correspondia a determinado artigo de uma enciclopédia virtual. Pouco tempo decorrido e

[2]Disponível no sítio da Wikipédia em `http://dumps.wikimedia.org/ptwiki/20110425/`.

surgiram espaços de nomes (*namespaces*) especiais, para guardar tipos de páginas que não correspondem a artigos. A secção 5.1.1 descreve um conjunto destes tipo de espaços. Enquanto que na navegação da Wikipédia é mais ou menos claro o que corresponde a um artigo da enciclopédia, e o que constitui um documento auxiliar de gestão, na cópia estática é necessário fazer essa divisão de forma manual, detetando em que espaço cada documento está.

Exemplos destes espaços de gestão são os *redireção* e *desambiguação*, que albergam páginas que servem de entradas preferenciais ou entradas de desambiguação para artigos (e que são descritos de seguida). Existe um outro espaço de gestão muito importante, denominado de *pré-definição*, que é explicado na secção 3.2.

3.1.1 Páginas de desambiguação

As páginas de desambiguação são usadas em situações em que uma palavra é polissémica. Nestes casos o utilizador é confrontado com um conjunto de resumos das páginas que representam cada um dos possíveis sentidos dessa palavra.

Por vezes a página de desambiguação não é logo apresentada. Por exemplo, ao procurar por *banco* o utilizador é redirecionado automaticamente para a página sobre a instituição financeira. Junto com o título da página aparece uma nota que permite ao utilizador saber que existem outros significados para a palavra, e deste modo aceder à página de desambiguação.

No entanto, se procurar por uma palavra ainda mais genérica, como *tipo*, a página de desambiguação é logo apresentada.

3.1.2 Redirecionamento

Durante a preparação da coleção do Págico foram encontrados dois tipos de redirecionamento, um dos quais está a cair em desuso.

O tipo de redirecionamento oficial serve para que um utilizador que procure um título que representa um tópico polimórfico (que pode ser descrito de diversas formas) o consiga encontrar. Exemplos são a pesquisa de um plural (*cavalos* em vez de *cavalo*) ou mesmo outro tipo de palavras relacionadas (*escravo* em vez de *escravidão*). Nestas situações a Wikipédia faz a ligação direta da pesquisa à página de destino, sem passar por uma página com o título procurado. No entanto, e junto do título (tal como no caso de palavras com página de desambiguação), é apresentada a forma original procurada pelo utilizador (*Escravidão (Redirecionado de Escravo)*).

O outro tipo de redirecionamento encontrado usa (ou usava) uma página intermédia, quase que como uma entrada remissiva num dicionário, que indicava ao utilizador que devia usar outra palavra para procurar a página desejada. Sendo apenas esta a informação que esta página continha não fazia sentido a sua existência, e talvez tenha sido essa a razão pela qual foram desaparecendo (durante a escrita deste artigo não se encontrou nenhum exemplo ilustrativo deste tipo de redirecionamento, no entanto foram encontrados vários casos na versão estática utilizada—que, note-se, tem cerca de um ano de idade).

3.2 A sintaxe MediaWiki

A sintaxe usada na Wikipédia é a sintaxe do sistema de Wiki MediaWiki. Não faz sentido nesta secção descrever toda a sintaxe suportada, já que corresponde a uma sintaxe Wiki comum, em que são usados carateres ASCII para a formatação do texto. A descrição oficial desta linguagem pode ser consultada, por exemplo, em `http://en.wikipedia.org/wiki/Wikipedia:Cheatsheet`.

Faz sentido, sim, referir o mecanismo de macros usado por esta linguagem, uma vez que se tornou uma pedra no processo de construção da coleção.

O mecanismo de macros permite que se definam abreviaturas, opcionalmente parametrizadas, que expandam em sintaxe Wiki ou diretamente em notação HTML.

Estas macros são definidas num espaço próprio (denominado *pré-definição* na Wikipédia portuguesa). Um exemplo de uma pré-definição é "`http://pt.wikipedia.org/wiki/Predefinição:POR`", que é uma macro para a inclusão da bandeira portuguesa juntamente com uma hiperligação para o artigo *Portugal*. Deste modo, basta usar `{{POR}}` numa página para que esta seja expandida na dita bandeira e hiperligação.

Existem macros bastante mais complexas. Um exemplo de uma macro parametrizada é a "`http://pt.wikipedia.org/wiki/Predefinição:Bandeira`," que permite a inclusão de bandeiras de qualquer país, com possibilidade de escolher uma variante (por exemplo, a da monarquia portuguesa), o tamanho da bandeira e o texto a ser apresentado. Um exemplo de uso desta macro será `{{Bandeira|Alemanha|império}}`.

Estas macros podem conter código condicional, opções condicionais, opções com valores por omissão e mais uma panóplia de opções que as tornam muito poderosas. Por exemplo, as

célebres tabelas (denominadas por *infobox*) usadas em páginas como as de países, cidades ou animais, que sistematizam alguma informação numa barra vertical ao lado direito, são geradas usando macros.

4 Construção da coleção do Págico

Tendo sido decidido que o formato original da Wikipédia não seria o ideal para a coleção, por obrigar os participantes a compreender o funcionamento quer da sintaxe Wiki, quer das macros, foi decidido que a melhor opção seria converter os artigos em documentos XHTML. É certo que podíamos ter optado por soluções como a apresentada em (Junior et al., 2011), em que a Wikipédia é, de algum modo, simplificada ou sumariada, mas passaríamos a estar mais longe do que é a Wikipédia original.

Em todo o caso, a escolha da conversão da Wikipédia em XHTML faz sentido uma vez que grande parte da recolha de informação dos dias que correm é feita sobre a Rede, em que grande parte dos documentos estão codificados em HTML ou XML, ou sobre documentos estruturados, armazenados por ferramentas específicas e que, na sua grande maioria, também são armazenadas em XML.

O uso de HTML (ou XHTML) como formato de eleição para a coleção do Págico teve outras vantagens, nomeadamente o de possibilitar o uso de uma ferramenta já desenvolvida para a gestão de avaliações deste tipo (o SIGA(Costa, Mota e Santos, 2012), por exemplo).

Nesta secção faremos uma apresentação inicial das alternativas para o processamento da coleção e geração de documentos XHTML, seguindo-se uma breve explicação de quais as ferramentas escolhidas, e de como foram usadas. Terminaremos com alguns dos problemas encontrados, bem como a solução adotada.

4.1 Ferramentas disponíveis

Grande parte das ferramentas disponíveis para a conversão da Wikipédia para outros formatos não tem tido atualizações recentemente[3]. Além disso, o facto de serem ferramentas não desenvolvidas pelos programadores da ferramenta MediaWiki leva a que não suportem a totalidade da sintaxe usada na Wikipédia. Ora, não havendo atualizações para estas ferramentas, e estando a Wikipédia em constante evolução, este problema é acentuado. Foram testadas várias ferramentas, como o *FlexBisonParse*, *Wiki2XML mediawiki-parser*, entre outros. Alguns não se conseguiram instalar, outros não reconheciam o formato XML da Wikipédia, e outros ainda geravam documentos de forma não satisfatória.

A abordagem mais prometedora seria a instalação de um servidor HTTP e uma base de dados para onde se importasse toda a Wikipédia, e instalar uma versão recente do MediaWiki. Tendo esta configuração, muitas ferramentas estavam disponíveis, e mesmo que não estivessem, uma ferramenta de *crawling* conseguiria, de forma simples, obter uma cópia local em HTML. No entanto a meta-informação (como quais as páginas que são de redireção) seria perdida.

A primeira ferramenta que mostrou resultados aceitáveis foi a `mwlib`[4], um conjunto de conversores em Python. Dada a proximidade do evento optou-se por usar esta biblioteca mesmo com todos os problemas encontrados (e que serão descritos mais à frente).

Para auxiliar o processo, foi usado um módulo Perl, `MediaWiki::DumpFile`[5], que permite percorrer a cópia estática em XML e extrair meta-informação.

4.2 Abordagem adotada

O processo detalhado de conversão do formato XML em ficheiros XHTML está descrito na página do Págico, em `http://linguateca.pt/Pagico/`. Nesta secção limitar-nos-emos a enumerar os passos necessários.

O processamento foi feito com base na cópia estática da Wikipédia, nomeadamente na sua cópia em formato XML, de nome `pages-articles.xml.bz2`. Este documento inclui todos os artigos da Wikipédia num único documento XML. A anotação XML é usada para toda a meta-informação, e os artigos estão descritos de forma textual, na sintaxe wiki.

Infelizmente a ferramenta que escolhemos (`mwlib`) foi desenvolvida para a versão inglesa da Wikipédia, o que nos trouxe alguns problemas. Nomeadamente, foi necessário realizar alterações diretamente no código fonte da ferramenta para que esta considerasse o documento XML na língua portuguesa.

O módulo Perl `MediaWiki::DumpFile` percorre todo o ficheiro XML obtendo meta-informação sobre cada artigo e, dependendo do seu tipo, tomando diferentes ações. No caso de

[3] Existe uma lista de ferramentas de conversão disponíveis em `http://www.mediawiki.org/wiki/Alternative_parsers`.

[4] Disponível em `http://pediapress.com/code/`.

[5] Disponível em `http://search.cpan.org/~triddle/MediaWiki-DumpFile-0.2.1/`.

ser um artigo comum, a ferramenta da `mwlib` para conversão em XML era invocada. No caso de ser uma página de redireção oficial, era gerado um documento HTML apenas com a ligação para a página oficial. Finalmente, em casos especiais, como páginas de desambiguação e páginas referentes a imagens, foram simplesmente descartadas.

Os documentos produzidos em XHTML foram arrumados numa árvore de diretorias, organizados pelos três primeiros carateres do título do documento. Além disso, os documentos foram processados pela ferramenta `xmllint` para garantir a correção dos documentos gerados.

4.3 Problemas encontrados

Foram vários os problemas encontrados durante a criação da coleção, o que explica a disponibilização quase consecutiva de 7 versões da coleção. Muitos destes problemas deveram-se a comportamentos não esperados por parte das ferramentas utilizadas. Por exemplo, a primeira versão disponibilizada a 1 de Agosto de 2011 incluía algumas páginas de redireção não detetadas.

Outras versões foram criadas por pequenos erros incluídos na preparação das coleções anteriores, como a incorreta normalização de títulos (carateres não previstos) ou a correção das hiperligações internas à coleção.

No entanto, os principais problemas encontrados foram as páginas de redireção não oficiais e o processamento das macros.

Em relação às páginas de redireção não oficiais, a decisão foi ignorar. Felizmente, não foram detetadas muitas destas páginas. Em todo o caso, a decisão seria a mesma, já que não existe uma forma clara para distinguir a página de redireção (intermédia) de uma página comum.

Processar as macros de forma satisfatória foi um problema mais complicado. Estas macros não podem ser ignoradas, já que levaria a que muita informação fosse perdida. Veja-se por exemplo a macro `{{POR}}` apresentada anteriormente, que se fosse ignorada levaria a que grande parte das ligações à página de Portugal fossem perdidas.

Embora os autores das `mwlib` digam que a ferramenta reconhece e trata corretamente as macros, não o conseguimos fazer para a versão portuguesa da Wikipédia (possivelmente pelo uso de *Predefinição* como prefixo, em vez do termo usado na Wikipédia inglesa, *Template*).

A solução foi implementada na casa: criou-se uma base de dados de macros, pré-processando o documento XML da Wikipédia, e para todas as páginas de pré-definição, foi introduzido um registo na base de dados, mapeamento do seu nome (nome da macro) e o conteúdo gerado pela macro (ignorando comentários usados para explicar como a macro se deve usar). Posteriormente, ao processar a Wikipédia, as macros seriam substituídas pela expansão respetiva.

Infelizmente esta abordagem não foi totalmente satisfatória, dado existir um conjunto de macros que geram etiquetas XHTML diretamente. Ora, ao interpolar as macros no XML com essas novas etiquetas, o documento XML deixava de ser bem formado, e a ferramenta `mwlib` não era capaz de o processar. Esta foi a principal razão pela qual se perderam as *Infoboxes* já mencionadas. Dada a necessidade de estabilizar rapidamente a coleção, e de estas caixas, embora contendo informação relevante, terem pouco que ver com língua natural (os dados são tabelados), a equipa do Págico decidiu ignorar este problema.

Existiu ainda um pequeno conjunto de macros que não foram expandidas corretamente dada a sua complexidade (número de argumentos, argumentos pré-definidos, aninhamento de macros, etc.).

5 Caracterização do Cartola

Esta secção faz uma caracterização preliminar do conteúdo do Cartola. Concretamente, apresenta estatísticas relativas à coleção do Págico, bem como diversas estatísticas relativas à subcoleção do monte do Págico. Esta subcoleção contém todos os documentos usados como resposta aos tópicos bem como os usados como justificações das respostas, não distinguindo se foram dados pelos criadores de tópicos ou pelos participantes.

O objetivo desta caraterização é permitir que o leitor consiga julgar a dificuldade (ou facilidade) da participação no Págico. Além disso, permite ter uma noção da abrangência dos tópicos em relação à coleção disponibilizada.

5.1 A coleção do Págico

Para que se tenha uma ideia do espaço de procura das páginas que podem ser respostas aos tópicos do Págico, apresentamos aqui várias quantificações em relação à coleção.

Começaremos por analisar o tamanho da coleção em número de documentos, e em número de documentos por tipo (ou *espaço de nomes*), o que indicará qual a percentagem de documentos da coleção que constituíam, realmente, espaço de procura das respostas.

Após a divisão de páginas pelo seu tipo, um sistema automático poderia tentar indexar os artigos pelas categorias que são usadas para os classificar. Deste modo, na secção 5.1.2 apresentamos algumas estatísticas que permitem analisar até que ponto as categorias usadas na Wikipédia podem ser úteis, ou não, na indexação dos artigos, e facilitação na pesquisa de respostas.

As secções que se lhe seguem tentam caracterizar a coleção de um ponto de vista mais concreto: qual é o tamanho da coleção? qual o número médio de palavras por artigo? Embora pouco relevante para a construção de um sistema ou para a indexação dos artigos, esta informação permite-nos saber o que constitui um artigo da coleção.

Finalmente, será apresentada uma análise temporal que permite caracterizar a coleção em termos de atualiadade. Possivelmente, esta análise é pouco relevante para o Págico, mas acaba por demonstrar que a maior parte dos artigos da Wikipédia portuguesa foram atualizados nos últimos 12 meses. Este facto só por si justifica a relevância em se ter criado uma nova coleção para o Págico (especialmente quando o Págico se propõe a sugerir temas ligados à cultura portuguesa), uma vez que a coleção do GikiCLEF foi criada a partir de uma versão de 2008 da Wikipédia.

5.1.1 Tipos de páginas

A coleção pode ser dividida em várias partições, de acordo com o tipo de conteúdo das páginas: páginas de pré-definições (com definições de funções, macros, etc.), páginas de desambiguação (que permitem ao utilizador escolher qual o artigo que realmente lhe interessa), páginas de redirecionamento (que funcionam como entradas remissivas), páginas relativas a conteúdo audiovisual (que descrevem imagens, sons, etc.) e as páginas de artigos propriamente ditos. A tabela 1 apresenta o número de páginas para cada um destes tipos. Destas, apenas as páginas relativas a conteúdo audiovisual não foram incluídas na coleção.

Tipo	Nº de documentos
Páginas de pré-definição	32 900
Páginas de desambiguação	5 006
Páginas de redireção	574 077
Páginas de audiovisuais	9 678
Artigos (e anexos)	856 005

Tabela 1: Distribuição de páginas da coleção por tipo.

Embora sejam 689 629 as páginas que fazem parte da coleção, e que não correspondem aos tipos descritas anteriormente, destes apenas 856 005 documentos correspondem a artigos propriamente ditos (e a anexos), onde, em princípio, se encontrarão as respostas aos tópicos do Págico.

Ou seja, uma quantidade razoável de documentos contidos na coleção não eram relevantes, nem constituíam o espaço de procura para as respostas aos tópicos do Págico. Uma nova versão da coleção poderia descartar essas páginas já que não traziam qualquer informação adicional, e acabam por gerar confusão, quer para os participantes, quer para os avaliadores.

5.1.2 Categorização das páginas

Um processo que pode ajudar na divisão do espaço de procura é o uso das categorias associadas a cada página da Wikipédia (colocadas em notação Wiki em cada página, na forma `[[Categoria:nome da categoria]]`). Estas categorias são colocadas de forma *ad-hoc* por quem contribui com artigos e, embora existam algumas regras definidas, não podem ser consideradas parte de uma estrutura classificativa estruturada, mas antes de, no melhor dos casos, uma estrutura classificativa de dois níveis. Na verdade, as estruturas classificativas mais próximas deste tipo de classificação são as *Folksonomy* (Sinclair e Cardew-Hall, 2008).

A demonstração desta anarquia é o número de categorias existente: 95 446 categorias para classificar 681 058 documentos (a diferença deste número de documentos para o número total de documentos — 689 829 — mostra a existência de mais de 8 500 artigos não categorizados), o que corresponde a uma média de 7 documentos por categoria. Também é relevante dizer que a página *Língua inglesa* (Wikipédia) é a que tem mais categorias associadas, num total de 62. Por sua vez, existem 32 652 categorias que contêm apenas uma página associada, e a categoria com mais páginas (32 645) corresponde aos *Asteroides da cintura principal*. As tabelas 2 e 3 resumem esta informação. Não são apresentados os respetivos histogramas na sua forma gráfica já que a discrepância de valores torna-os pouco legíveis.

Para facilitar a comparação com a caraterização da coleção composta apenas pelas páginas correspondentes a tópicos (secção 5.2), e dado que a maioria dos documentos tem entre 0 a 8 categorias associadas, a figura 1 apresenta uma estatística mais fina correspondente a este intervalo.

nº de documentos	total de cat.	percentual
]0, 1]	32 652	34.21%
]1, 66]	59 775	62.63%
]66, 130]	1 789	1.87%
]130, 194]	507	0.53%
]194, 260]	231	0.24%
]260, 345]	166	0.17%
]345, 442]	108	0.11%
]442, 592]	84	0.09%
]592, 862]	68	0.07%
]862, ∞[	65	0.07%

Tabela 2: Número de documentos por quantidade de categorias (p.ex. existem 32 652 categorias que só classificam um documento; e existem 65 categorias que classificam mais de 862 documentos).

nº categorias	total docs.	percentual
0	8 771	1.271%
]0, 8]	676 705	98.097%
]8, 15]	4 008	0.581%
]15, 23]	314	0.046%
]23, 33]	25	0.004%
]33, ∞[	6	0.001%

Tabela 3: Número de categorias por quantidade de documentos (p.ex, existem 8 771 documentos sem categorias associadas, e existem 6 documentos com mais de 33 categorias associadas).

5.1.3 Tamanho das páginas

O tamanho médio (incluindo toda a anotação wiki) destes artigos é de 3 169 bytes, cerca de 968 formas[6] (os artigos mais pequenos estão (ou estavam) vazios; o maior artigo, com o título *Anexo: Lista de espécies da família Salticidae* (Wikipédia)[7] tem 334 083 bytes (106 140 formas)).

[6]De realçar que os valores de formas aqui apresentados não correspondem a palavras uma vez que devido à grande quantidade de anotação Wiki presente nos documentos, apenas uma percentagem corresponde, realmente, a palavras. Além do mais, esta percentagem não é mantida entre páginas já que algumas (como a que é referida, com 106 140 formas) são tabelas com uma grande quantidade de anotação, e outras páginas, de artigos convencionais, têm uma quantidade de anotação bastante menor.

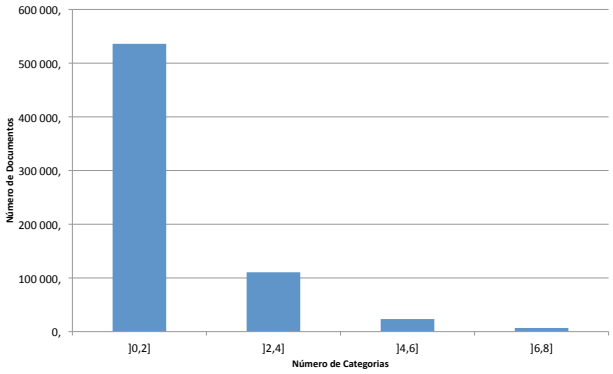

Figura 1: Número de categorias por quantidade de documentos, no intervalo de]0, 8] categorias.

nº de formas	nº docs	percentual
]0, 5]	1	0.00%
]5, 1042[	541 628	78.54%
]1042, 2075[	87 789	12.73%
]2075, 3108[	26 527	3.85%
]3108, 4141[	11 931	1.73%
]4141, 5176[	6 501	0.94%
]5176, 6232[	3 946	0.57%
]6232, 7378[	2 711	0.39%
]7378, 8707[	1 989	0.29%
]8707, 10256[	1 691	0.25%
]10256, 12439[	1 447	0.21%
]12439, 15585[	1 256	0.18%
]15585, 21968[	1 139	0.17%
]21968, ∞]	1 063	0.15%

Tabela 4: Número de documentos por classes de tamanhos (Por exemplo, a maioria dos documentos (78%) tem menos de 1042 formas).

5.1.4 Atualidade da coleção

O gráfico da figura 2, correspondente à tabela 5 mostra a evolução das páginas da coleção de acordo com a sua última edição.

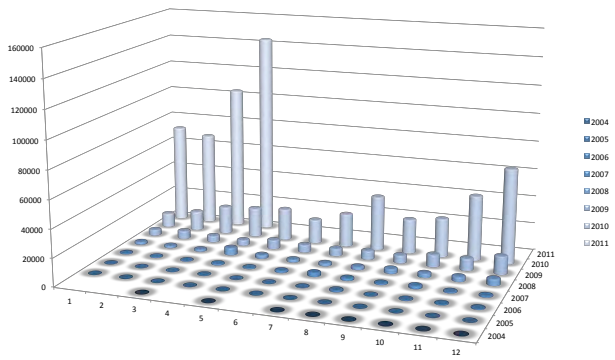

Figura 2: Número de artigos por ano/mês.

Embora o gráfico não permita ver as diferenças relativas aos primeiros anos torna mais visual a discrepância no número de artigos atualizados recentemente. Na verdade, esse valor aumenta à medida que nos aproximamos da atu-

[7]Note que este é o artigo maior em termos absolutos e não em termos de formas. Nesse caso, o artigo *Torneio de Wimbledon* (Wikipédia) estaria no topo, com 158 128 formas.

Ano	Jan.	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.	Total
2004			4		9		5	5	4	5	7	8	47
2005	9	3	17	16	74	61	33	30	64	16	39	25	387
2006	120	96	101	316	125	228	268	1329	271	528	638	726	4746
2007	681	590	487	1023	834	1461	2933	1760	1007	2199	970	1058	15003
2008	1977	1654	1554	5385	2812	2125	2123	2328	3570	3148	3574	4883	35133
2009	4330	5876	4665	4024	6559	5558	5369	6364	5804	8866	8768	13098	79281
2010	10131	13988	19879	21241	22941	17257	23927	39281	24860	27785	46672	68136	336098
2011	71369	67126	103464	143351									385310

Tabela 5: Número de artigos por ano/mês.

alidade, o que sugere uma atualização contínua dos conteúdos.

5.2 A subcoleção do monte do Págico

Nesta subsecção, vamos debruçar-nos sobre a subcoleção do monte do Págico, ou seja, o subconjunto da coleção constituído pelos documentos usados como resposta ou justificação pelos criadores dos tópicos, no processo de criação dos mesmos, e por todos os participantes no Págico (tanto sistemas automáticos como participações humanas). Por simplificação, usaremos o termo *documento de resposta*, independentemente desse documento ter sido usado como resposta ou justificação.

Primeiro faremos uma análise sem ter em conta se as respostas do monte estavam ou não corretas, e sem seguida teremos apenas em consideração os documentos de resposta que correspondem a respostas e justificações corretas.

5.2.1 Visão sobre todas as respostas

A figura 3 apresenta uma panorâmica sobre a distribuição do número de documentos de resposta determinados pelos criadores dos tópicos e encontrados pelos participantes no Págico. Como se pode constatar, para a maior parte dos tópicos, o número de documentos associados varia entre 175 e 250 documentos. Se nos restringirmos aos documentos que existem apenas na Wikipédia portuguesa, portanto sem equivalentes noutras línguas, então obtemos o gráfico da figura 4, onde se pode ver que, para a maior parte dos tópicos, entre 20% e 50% dos documentos de resposta existem unicamente na Wikipédia em português.

Os tópicos mais especificamente lusófonos, se assim considerarmos aqueles para os quais uma maior percentagem dos documentos de resposta existe apenas na Wikipédia em português, são sobre samba (**tópico 36** [Escolas de samba fundadas ou sediadas em morros cariocas.], **tópico 51** [Além do samba, que outros gêneros musicais são populares no carnaval brasileiro] e **tópico 86** [Compositoras brasileiras de samba]) e São

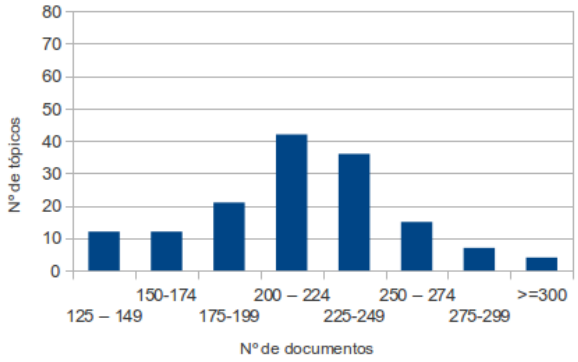

Figura 3: Número de tópicos agrupados por número de documentos de resposta.

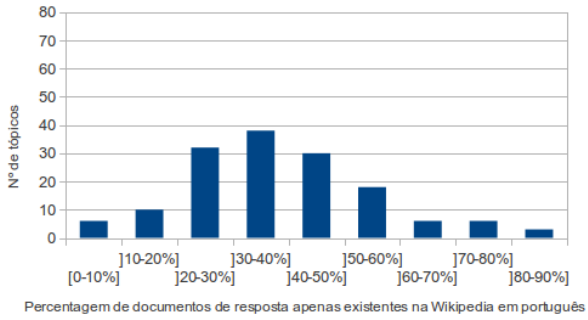

Figura 4: Número de tópicos agrupados pela percentagem de documentos de resposta apenas existentes na Wikipédia em português.

Tomé e Príncipe (**tópico 131** [Quem descobriu São Tomé e Príncipe?] e **tópico 95** [Partidos políticos de São Tomé e Príncipe]).

No pólo oposto, os tópicos para os quais uma menor percentagem dos documentos de resposta existe apenas na Wikipédia em português, os tópicos sobre desporto estão bem representados (**tópico 58** [Países que venceram a Copa do Mundo em uma disputa de pênaltis], **tópico 137** [Eventos onde Maria de Lurdes Mutola foi medalha de ouro] e **tópico 39** [Modalidades esportivas em que países lusófonos já ganharam medalha de ouro nos Jogos Olímpicos.]).

A figura 5 mostra o número total de palavras dos documentos de resposta. Este número varia bastante de tópico para tópico, havendo tópicos

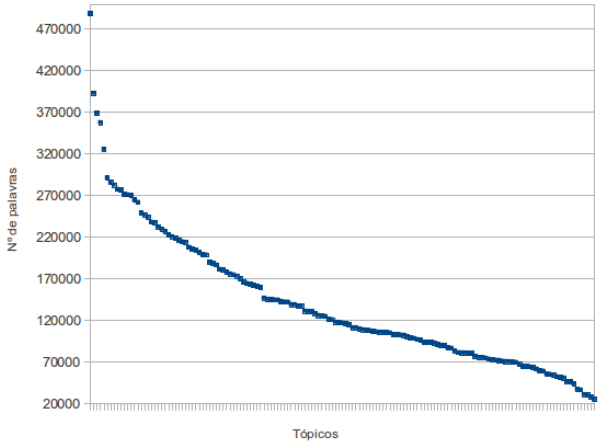

Figura 5: Número de palavras dos documentos de resposta.

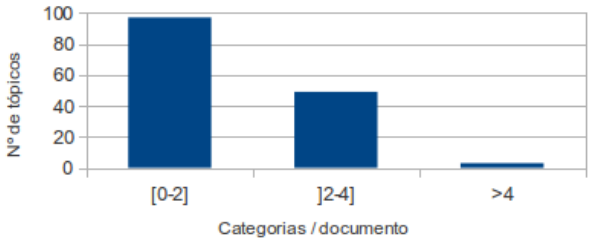

Figura 6: Número de tópicos agrupados pelo número de categorias em que estão classificados os documentos de resposta.

com menos de 50000 palavras, enquanto outros têm mais de 300000 palavras.

A figura 6 ilustra a distribuição do número de categorias por documento em que estão classificados os documentos de resposta. Como se pode constatar para a maior parte dos tópicos, este número não ultrapassa as duas categorias por documento.

A tabela 6 apresenta os cinco tópicos com maior e menor número de documentos de resposta. É curioso verificar que os cinco tópicos para os quais foram encontrados menos documentos de resposta são todos sobre temas africanos o que parece indicar que a Wikipédia conterá menos informação sobre esses temas.

5.2.2 Visão sobre as respostas corretas do Págico

A figura 7 apresenta uma panorâmica sobre a distribuição do número de documentos de resposta corretos, ou seja, relativos às respostas e justificações determinadas pelos criadores dos tópicos e encontradas pelos participantes no Págico que foram consideradas corretas. Como se pode constatar, para a maior parte dos tópicos este número situou-se abaixo dos dez documentos. Se nos restringirmos aos documentos que existem apenas na Wikipédia portuguesa, portanto sem equivalentes noutras línguas, então obtemos o gráfico da figura 8. Estes valores diferem bastante dos encontrados para todos os documentos de resposta (cf. figura 4). Neste caso para um terço dos tópicos, a percentagem de documentos de resposta apenas na Wikipédia em português situa-se entre os 0% e 10%, existindo apenas dois tópicos onde este valor é superior a 90% (**tópico 41** [Congressos ou conferências que têm por tema as relações culturais e/ou sociais entre África e demais países lusófonos] e **tópico 54** [Igrejas do Rio de Janeiro construídas por irmandades ou confrarias de negros]).

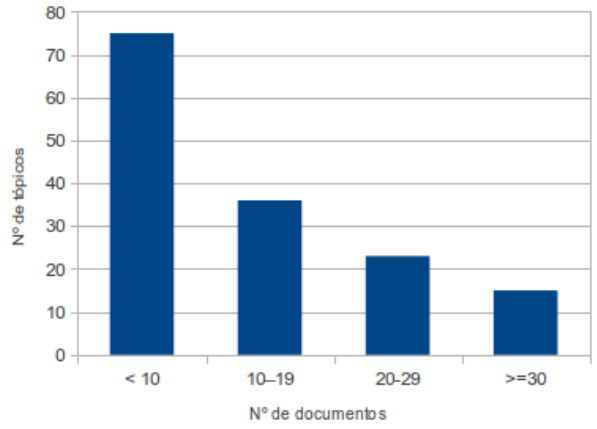

Figura 7: Número de tópicos agrupados por número de documentos de resposta corretos.

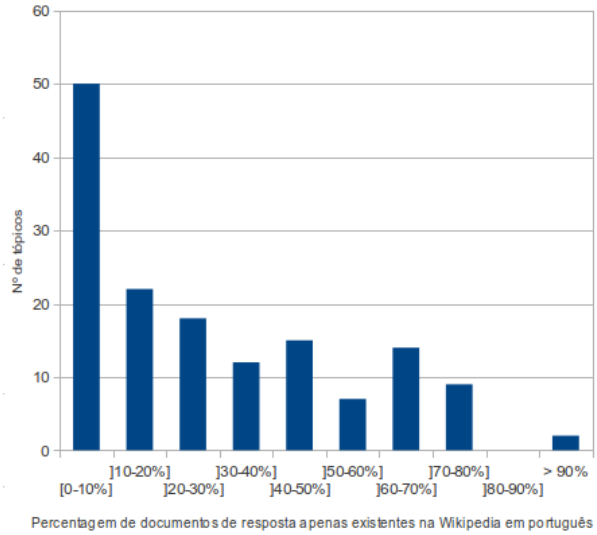

Figura 8: Número de tópicos agrupados pela percentagem de documentos de resposta corretos apenas existentes na Wikipédia em português.

A figura 9 mostra o número total de palavras dos documentos de resposta correspondentes a respostas e justificações corretas. Este número

ID	Tópico	# Documentos
83	Que equipes da primeira divisão do futebol brasileiro desceram para a segunda divisão e nunca mais conseguiram voltar?	330
142	Locais referidos n' "Os Lusíadas"	327
17	Documentários sobre políticos brasileiros.	325
29	Escritores lusófonos que se filiaram a partidos políticos	315
35	Que autores não lusófonos escreveram sobre o Brasil nos séculos XVIII e XIX?	294
	(...)	
109	Candidatos a alguma das eleições presidenciais na Guiné-Bissau	129
95	Partidos políticos de São Tomé e Príncipe	128
129	Antigos alunos da Universidade Eduardo Mondlane e da sua antecessora, a Universidade de Lourenço Marques	128
100	Ilhas de Moçambique	125
121	Frutos de Angola	125

Tabela 6: Tópicos com maior e menor número de documentos de resposta.

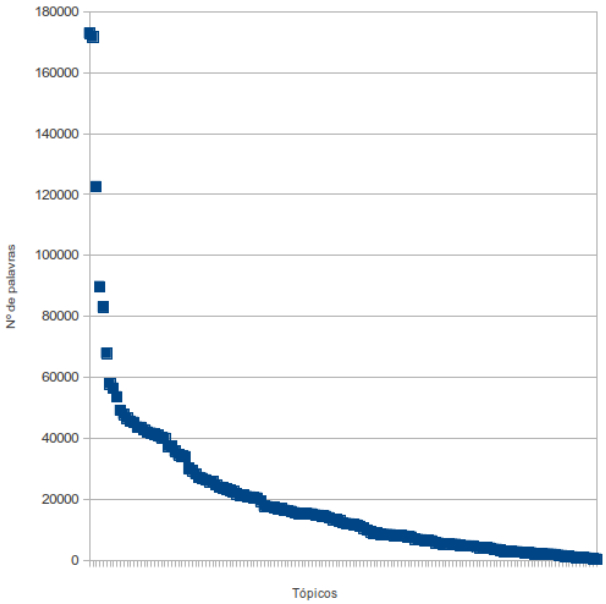

Figura 9: Número de palavras dos documentos de resposta corretos.

varia bastante de tópico para tópico, havendo tópicos com menos de um milhar de palavras, enquanto outros têm mais de cem mil palavras.

A figura 10 ilustra a distribuição do número de categorias por documento em que estão classificados os documentos de resposta corretos. Como se pode constatar para a maior parte dos tópicos, o número de categorias por documento situa-se entre as zero e as quatro categorias.

A tabela 7 apresenta os cinco tópicos para os quais foram determinados o maior e menor número de documentos de resposta corretos. Em relação aos tópicos com menos documentos de resposta corretos, a maior parte deles são sobre temas africanos, tal como se verificou considerando o conjunto total de respostas (corretas e incorretas). Relativamente aos tópicos com mais respostas corretas, parecem ser tópicos que de facto têm naturalmente um número elevado de respostas tais como **tópico 19** [Tribos indígenas que vivem na Amazônia] e **tópico 147** [Museus em capitais de países lusófonos].

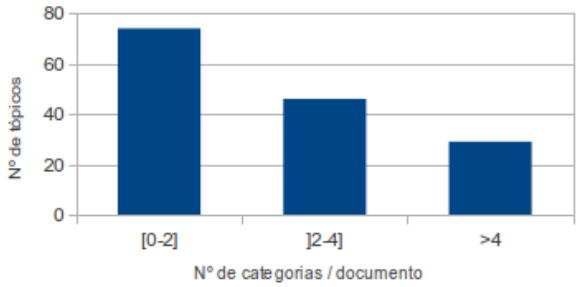

Figura 10: Número de tópicos agrupados pelo número de categorias por documento em que estão classificados os documentos de resposta corretos.

6 Comentários finais

É certo que a coleção desta edição do Págico tem muitos problemas. O principal problema é depender de uma ferramenta externa para a produção dos documentos num formato menos complicado. Poder-se-ia ter disponibilizado aos participantes a versão original em XML disponibilizada pela própria Wikipédia, mas isso obrigaria os participantes a processar a marcação Wiki, processamento este que iria influenciar os resultados da participação, mas que nada têm a ver com a tarefa do Págico de encontrar as respostas aos tópicos.

ID	Tópico	# Documentos
19	Tribos indígenas que vivem na Amazônia.	95
147	Museus em capitais de países lusófonos	62
144	Locais referidos n' "Os Lusíadas"	51
79	Povos indígenas brasileiros considerados extintos.	50
106	Vice-reis da Índia Portuguesa	48
	(...)	
110	Políticos da África lusófona que estudaram na União Soviética	2
54	Igrejas do Rio de Janeiro construídas por irmandades ou confrarias de negros.	1
132	Deputados da FRELIMO	1
116	Escritores moçambicanos que receberam o Prémio Camões	1
55	Escritores estrangeiros que visitaram Portugal no século XIX e que publicaram descrições das suas viagens	1

Tabela 7: Tópicos com maior e menor número de documentos de resposta corretos.

Numa próxima edição a solução deverá passar por usar uma versão do motor da Wikipédia em modo local, e pela extração dos documentos HTML através de *crawling*. Esta abordagem irá desencadear um conjunto de outros problemas mas que, esperamos, serão menos graves que os encontrados com a coleção atual.

Com a compilação do Cartola, o recurso público criado no decurso do Págico, pretendemos que o trabalho e a experiência no Págico possa ser o mais proveitosa possível, mesmo após o término do mesmo. Ou seja assumindo naturalmente que nem sempre tomámos as melhores opções no decorrer da organização do Págico, disponibilizamos todos os resultados obtidos, para que possam ser usados e eventualmente melhorados por quem estiver interessado nas áreas abordadas pelo Págico.

Ideias para trabalho futuro seriam, por exemplo:

- o estudo da evolução da Wikipédia ao longo dos últimos anos, usando para isso quer as coleções desenvolvidas no contexto do GikiCLEF e no contexto do Págico, ou diretamente usando as cópias estáticas disponibilizadas pela Wikipédia.
- aferir a lusofonia da Wikipédia portuguesa, por um lado, a nível de conteúdo, por exemplo, analisando os topónimos e gentílicos usados nas categorias das páginas, e, pelo outro, em termos de quem a escreve, por exemplo, analisando a grafia e o vocabulário.

Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN, e em 2011 pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN). O trabalho de Alberto Simões foi parcialmente suportado pela bolsa da Fundação para a Ciência e a Tecnologia SFRH/BPD/73011/2010.

Agradecemos a Cláudia Freitas e Alice Gonçalves pela paciência de nos irem relatando os vários erros encontrados na coleção do Págico enquanto utilizadoras da mesa no SIGA, o que ajudou a melhorar a qualidade do recurso criado.

Estamos também gratos a Sandra Aluísio, Diana Santos e António Teixeira pelos comentários e sugestões que recebemos durante a preparação do artigo e que enriqueceram o mesmo, tornando-o também mais claro.

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.

Freitas, Cláudia. 2012. A lusofonia na wikipédia em 150 tópicos. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Freitas, Cláudia, Paulo Rocha, Cristina Mota, Luís Costa, e Diana Santos. 2012. O que é uma resposta? Notas de uns avaliadores esta-

fados. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Junior, Arnaldo Candido, Ann Copestake, Lucia Specia, e Sandra Maria Aluísio. 2011. Towards an on-demand simple portuguese wikipedia. Em *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, pp. 137–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mota, Cristina. 2012. Resultados págicos: participação, medidas e pontuação. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Santos, Diana, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Julia Schulz, Yvonne Skalban, e Erik Tjong Kim Sang. 2010. Gikiclef: Crosscultural issues in multilingual information access. Em Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may, 2010. European Language Resources Association (ELRA).

Sinclair, James e Michael Cardew-Hall. 2008. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, February, 2008.

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 `SV`s 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.

Referncias

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.

Medindo o precipício semântico

Nuno Cardoso
Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
ncardoso@xldb.di.fc.ul.pt

Resumo

Este artigo descreve a minha participação na avaliação conjunta Págico e detalha a estratégia seguida para a participação, que usou um sistema de recuperação de informação geográfica com módulos especializados na interpretação e reformulação semântica de consultas. A estratégia seguida revelou-se demasiado complexa para gerar saídas automáticas, e a participação resumiu-se no envio de três saídas básicas. O artigo faz um resumo das lições aprendidas e tece algumas considerações sobre trabalho futuro.

Palavras chave

Págico, avaliação conjunta, recuperação de informação geográfica, reformulação semântica de consultas

1 Introdução

A minha participação no Págico foi feita com um sistema de recuperação de informação geográfica (RIG) desenvolvido no âmbito do projecto GREASE (Silva et al., 2006) e da Linguateca (Santos et al., 2004). O projecto GREASE investigou estratégias de adição de raciocínio geográfico em sistemas de recuperação de informação, para que estes obtenham um melhor desempenho na recuperação de documentos para consultas com âmbito geográfico.

Este protótipo RIG participou em várias tarefas de avaliação do GeoCLEF (Gey et al., 2007), com o intuito de avaliar estratégias de combinação de relevância geográfica e relevância textual em consultas típicas de recuperação de informação (Cardoso et al., 2008; Cardoso e Santos, 2008). Contudo, as suas capacidades semânticas ainda eram limitadas, resumindo-se à captura simples de entidades geográficas nas consultas e nos documentos.

Já no âmbito do meu trabalho de doutoramento, o Renoir foi criado e desenvolvido para participar em avaliações conjuntas piloto como o GikiP (Santos et al., 2009), GikiCLEF (Santos et al., 2010) e NTCIR (Cardoso e Silva, 2010a), onde há um maior foco na interpretação das consultas e no raciocínio de respostas.

O Renoir é um módulo de reformulação de consultas que procura compreender e enriquecer as consultas com informação semântica, com o objectivo de tornar a intenção do utilizador mais clara, e melhorando a capacidade do sistema de RIG em recuperar documentos relevantes.

Avaliações como o Págico são de grande relevância para a avaliação de sistemas de RIG, uma vez que é necessário que os sistemas RIG tenham uma capacidade robusta de interpretação de intenções dos utilizadores (como é o caso da deteção de âmbitos geográficos), e porque existe uma forte presença de desafios geográficos na tarefa (Cardoso, 2008a; Santos, Cardoso e Cabral, 2010).

O meus objectivos na participação no Págico são dois: i) medir o desempenho do Renoir na interpretação dos tópicos e na geração de respostas correctas, e ii) medir o desempenho do sistema de RIG na recuperação de documentos que correspondam a entidades que são respostas correctas.

No entanto, as saídas oficiais enviadas foram obtidas apenas com o sistema de RIG a funcionar sem o auxílio do Renoir, e em configurações básicas. Apesar da estratégia proposta implicar grandes dificuldades na geração de resultados de forma automática, acredito que um sistema que se proponha realizar uma tarefa como a apresentada pelo Págico, terá de seguir uma estratégia semelhante à que segui, e que passo a explicar de seguida.

2 Estratégia semântica proposta

O tipo de tópicos usado no Págico e em avaliações similares, embora sejam avaliações que procurem fundir a recuperação de informação (RI) com a resposta a perguntas (RAP), são na sua grande maioria tópicos de RAP, ou seja, perguntas que requerem como resposta uma ou mais entidades (cidades, pessoas, organizações, etc.), coadjuvados se possível com a justificação

para a escolha da(s) resposta(s).

Em avaliações conjuntas para sistemas de RI, os tópicos procuram imitar as necessidades de informação típicas dos utilizadores (Peters e Braschler, 2001; Rachel Aires et al., 2003; Voorhees e Harman, 2005). Normalmente, os tópicos referem um determinado tema de interesse, e espera-se que o sistema de RI encontre documentos relevantes para esse mesmo tema. No caso de avaliações de sistemas RIG, os tópicos incluem um determinado âmbito geográfico. Na prática, os temas escolhidos para os tópicos são fortemente condicionados pelas colecções usadas, uma vez que é preciso existir uma quantidade mínima de documentos relevantes para que o tópico possa ser considerado útil para a avaliação.

Assim sendo, todo o sistema que se proponha participar em avaliações como o Págico precisa de ter uma noção precisa do que são entidades, como encontrar essas entidades e atribuir classificações semânticas, compreender o papel dessas entidades no contexto da pergunta, e ter a capacidade de calcular a probabilidade de que um outro conjunto de entidades são respostas adequadas para a pergunta inicial.

Por outras palavras, um sistema de RI não tem essa capacidade; Um sistema de RIG também não tem, apesar da sua necessidade de detectar entidades geográficas e de e de raciocinar sobre o domínio geográfico. Como tal, a primeira tarefa é dotar o sistema de RIG da capacidade de analisar as coleções e consultas a uma maior profundidade semântica. Esse papel é desempenhado pelo Renoir, que se encontra esquematizado na Figura 1.

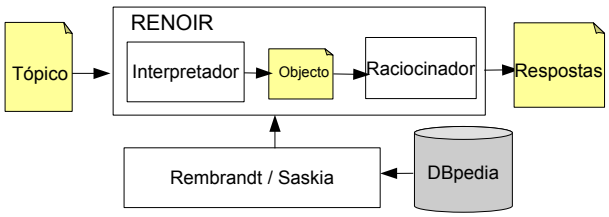

Figura 1: Esquema de funcionamento do reformulador de consultas Renoir.

O funcionamento do Renoir é resumido da seguinte forma: o tópico é analisado pelo *interpretador*, que converte as perguntas em *objectos* computacionais, que procuram representar as diversas propriedades da pergunta de uma forma simples e inteligível para que o *raciocinador* possa usar. O *raciocinador* decide qual a melhor estratégia para obter respostas correctas e, se possível, gera essa lista de entidades candidatas a respostas finais.

Para a obtenção de respostas, o raciocinador recorre frequentemente à DBpédia, uma base de dados gerada a partir de instantâneos da Wikipédia (Auer et al., 2007). A DBpédia pode ser acedida em `http://dbpedia.org`, e permite a consulta dos seus dados usando consultas SPARQL[1]. Em resumo, a função do raciocinador é a de traduzir a pergunta inicial, formulada em língua natural, num conjunto de perguntas SPARQL, e verificar se as respostas obtidas correspondem ao tipo de entidades esperadas pela pergunta.

Um exemplo de funcionamento do Renoir pode ser dado pelo tópico 81, "Igrejas em Macau", que o raciocinador do Renoir pode converter na seguinte consulta SPARQL:

```
SELECT ?Churches WHERE {
 ?Churches skos:subject
 <http://dbpedia.org/resource/Category:
  Churches_in_Macau>
}
```

Na DBpédia 3.7, a consulta gera os seguintes 6 resultados:

```
dbpedia.org/resource/Ji_Dou_Church
dbpedia.org/resource/St._Dominic%27s_Church_%28Macao%29
dbpedia.org/resource/St._Joseph%27s_Seminary_and_Church
dbpedia.org/resource/Ruins_of_St._Paul%27s
dbpedia.org/resource/Macau_Protestant_Chapel
dbpedia.org/resource/St._Lazarus%27_Church
```

Vamos analisar ao detalhe os componentes do Renoir.

2.1 Interpretador

O interpretador converte a pergunta, em linguagem natural, num objecto que represente essa pergunta de uma forma facilmente manipulável pelo programa. No processo, o interpretador procura identificar entidades mencionadas, expressões de pergunta (quantos, quais, etc) e outros padrões que possam ser mapeados a entidades na DBpédia.

O objecto é composto pelos seguintes elementos:

Tema, a entidade que define o tipo de resposta esperada. O tema pode ser mapeado como i) um recurso da DBpédia com uma propriedade `rdf:type` para um valor `skos:Concept`, ii) uma classe ontológica da DBpédia, ou iii) uma classificação semântica definida pelo HAREM (Santos e Cardoso, 2007; Santos et al., 2008), nesta ordem preferencial.

[1] `http://www.w3.org/TR/rdf-sparql-query/`

Condições , ou uma lista de critérios que filtram a lista de respostas candidatas. Cada condição é composta por i) uma propriedade ontológica da DBpédia, ii) um operador e iii) um recurso da DBpédia.

Tipo de Resposta Esperada (TRE), que define as propriedades que a lista final de respostas tem que ter.

O objecto é gerado mediante a aplicação de um conjunto de padrões sobre o tópico, previamente anotado pelo analisador morfossintático Palavras (Bick, 2000). Exemplificando com outro tópico do págico, #68: "Bandas brasileiras de punk formadas até 1980":

Mapear o tema: O primeiro conjunto de padrões detecta os termos que definem o tema. No tópico exemplo, a regra "`(Que)? <[nome]+ [adjectivo]*>`" captura os termos "bandas brasileiras", que foram previamente anotados como nome e adjectivos (a presença do termo "Que" é facultativo nesta regra). De seguida, estes termos são mapeados para o recurso da DBpédia `http://dbpedia.org/resource/Category:Brazilian_bands`, um recurso criado a partir da respectiva página da Wikipédia para essa categoria (e que, como tal, possui a propriedade `rdf:type` com o valor `skos:Concept`).

Mapear o TRE: Depois do tema mapeado, outros padrões determinam o TRE a partir do tipo de pergunta, e o tipo de tema. Para perguntas "Que X" tal como no tópico exemplo, o TRE é atribuído ao tema, ou seja, as respostas têm que ter uma propriedade `skos:subject` com o valor igual ao tema, o que significa que a resposta esperada tem de ser forçosamente uma banda brasileira. Note que, por exemplo, o padrão "Quantos X" atribui o TRE a um número, e diz ao raciocinador que a resposta final tem de ser o tamanho da lista de respostas, ou um valor obtido a partir de uma propriedade DBpédia.

Se o tema não for mapeado a um recurso DBpédia, é então mapeado a uma classe na ontologia DBpédia ou a uma categoria do HAREM, que também pode ser mapeado ao TRE. Supondo que o interpretador não consegue mapear "bandas brasileiras" a um recurso da DBpédia; usando um almanaque interno, o termo "Bandas" faz com que a TRE seja mapeada à classe `http://dbpedia.org/ontology/Band`. Por último, se o interpretador não consegue mapear a uma classe da ontologia DBpédia, é usada a categoria/tipo do HAREM `PESSOA/GRUPOIND`.

Mapear restrições: no tópico exemplo, há duas condições: a primeira filtra as respostas correctas a partir de uma lista inicial de bandas a aquelas que se formaram no Rio de Janeiro, e a segunda filtra as respostas para bandas formadas antes de 1980. O interpretador do Renoir deve ter um padrão que captura a expressão "formadas? [em|desde|até] X em Y", que gera duas condições: i) condição formada pela propriedade `dbpedia-owl:yearsActive`, operador `BEFORE`, e um valor 1980-01-01 (data), e ii) condição formada pela propriedade `dbpedia-owl:hometown`, operador `IS`, e uma entidade referente `http://dbpedia.org/resource/SÃčo_Paulo`.

2.2 Raciocinador

Dependo das propriedades encontradas no objecto, o raciocinador decide qual é a melhor estratégia para obter as respostas. A ação do raciocinador é uma lista de consultas SPARQL realizadas à DBpédia, para obter as respostas e justificações.

No exemplo dado, o objecto ideal teria uma TRE mapeada ao recurso DBpédia que melhore descreve o tipo de entidades que queremos obter como respostas, tal como `Category:Brazilian_punk_rock_groups`. As condições seriam a restrição do local de formação a São Paulo, e a data de formação anterior a 1980. O resultado ideal do raciocinador seria a seguinte consulta SPARQL:

```
SELECT ?x WHERE {
 ?x dcterms:subject <http://dbpedia.org/
resource/Category:Brazilian_punk_rock_groups> .
 ?x dbpedia-owl:hometown
 <http://dbpedia.org/resource/São_Paulo> .
 ?x dbpprop:yearsActive ?y .
  FILTER (?y < "1980-01-01"^^xsd:date) .
}
```

Com a DBpédia 3.7, esta consulta devolve 0 resultados. Relaxando a restrição do ano da formação, a DBpédia devolve dois resultados, `dbpedia.org/resource/Titãs` e `dbpedia.org/resource/Kleiderman`. Retirando a restrição do local de formação, a DBpédia devolve 10 resultados, o que significa que só 10 bandas são classificadas como bandas de punk brasileiras.

3 Integração do RENOIR no sistema de RIG

Como foi exemplificado na secção anterior, a ação do Renoir pode resultar frequentemente na

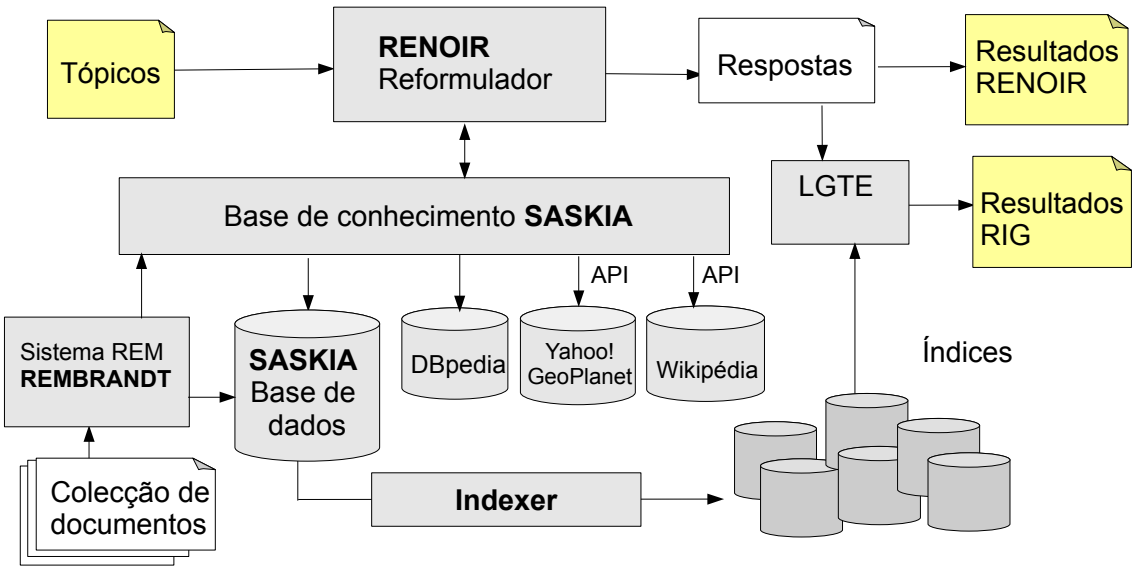

Figura 2: sistema de RIG usado para gerar respostas no Págico.

geração de nenhuma resposta. Existem muitas razões para tal:

1. O Renoir não possui padrões suficientes para que a interpretação das perguntas seja razoavelmente abrangente. Este é o principal desafio do Renoir, que precisa de ser mais tolerante a diferentes formulações de um tipo de conceito;

2. A dificuldade em gerar objectos que representem adequadamente a pergunta inicial. Um exemplo é o tópico 148, "Primeiras universidades de cada país lusófono". O conceito de universidades primogénitas é difícil de transcrever pelo Renoir; adicionalmente, o conceito de país lusófono é um conceito que requer um mecanismo de expansão para o formato interno mais conveniente, que é a lista explícita dos países abrangidos.

3. Mesmo assumindo que o objecto está bem descrito, é difícil decidir pelo conjunto de consultas SPARQL certas para a geração de respostas. No caso anterior do tópico #68, o raciocinador poderia optar por gerar consultas SPARQL que usasse uma categoria mais genérica, como por exemplo `Brazilian_music_groups`, e adicionar a restrição:

   ```
   ?x dbpedia-owl:genre
   <http://dbpedia.org/resource/Punk_Rock> .
   ```

   Contudo, esta estratégia iria obter nenhum resultado.

4. Quando o raciocinador obtém nenhum resultado, torna-se difícil de identificar o que está a provocar tal resultado. Além dos problemas de formulação das consultas SPARQL, há sempre a incerteza se a consulta é demasiado específica. Por exemplo, a informação do ano de formação de bandas pode ser uma informação rara nas páginas da Wikipédia, o que faz que a restrição temporal usada elimine respostas que possam estar correctas só porque a informação não existe por enquanto na DBpédia.

Por outras palavras, em teoria o Renoir não precisaria do sistema de RIG para gerar respostas válidas; na prática, o Renoir é demasiado frágil e muito pouco abrangente para gerar respostas de forma consistente, e é necessário recorrer a um sistema de RI/RIG.

Como tal, a participação no Págico depende mais da capacidade do sistema de RIG em recuperar documentos, do que da capacidade do Renoir em raciocinar sobre respostas. No entanto, o sistema de RIG usado pode recuperar documentos usando não só a similaridade textual, mas também similaridade em relação a entidades.

O Renoir consegue reconhecer TRE e entidades nas consultas, e se a coleção de documentos for previamente anotada e as entidades mencionadas devidamente reconhecidas e indexadas, é possível então adicionar uma camada semântica na geração das saídas.

No caso do tópico #120, "Cervejas consumidas em Angola", o sistema de RIG consegue medir a similaridade geográfica entre o âmbito da consulta (Angola) e os documentos que referem locais em Angola. Por outras palavras, um

	1000 respostas			100 respostas		
	#1	#2	#3	#1	#2	#3
Respostas	15000	15000	15000	1500	1500	1500
correctas, justificadas	436	329	398	129	94	102
correctas, não justificadas	38	25	29	9	11	9
Precisão	0,0291	0,0219	0,0265	0,0860	0,0627	0,0680
Pseudo-abrangência	0,1939	0,1463	0,1770	0,0574	0,0418	0,0454
Pseudo-medida-F	0,0506	0,0381	0,0461	0,0688	0,0501	0,0544
Precisão tolerante	0,0316	0,0236	0,0285	0,0920	0,0700	0,0740
Originalidade	120	151	54	0	0	0
Criatividade	634	514	517	93	83	68
Pontuação final	12,67	7,2	10,6	11,09	5,9	6,9

Tabela 1: Resultados das corridas com base na avaliação de 1000 respostas por tópico, ou de somente 100 respostas por tópico

documento que refere consumo de cerveja em Luanda, tem uma forte possibilidade de ser recuperado, uma vez que o sistema de RIG consegue determinar que Luanda está dentro do âmbito geográfico da consulta, da forma explicada ao detalhe em (Cardoso e Silva, 2010b).

A Figura 2 apresenta o sistema de RIG usado na geração de resultados para o Págico. O sistema baseia-se no modelo clássico de RI, onde os termos dos documentos (ou apenas os seus lemas) são indexados, e a recuperação de documentos faz-se com base no algoritmo BM25 para calcular a similaridade entre os termos dos documentos e os termos da consulta (Robertson et al., 1992). Na participação do Págico, os parâmetros k_1=1.2 e b=0.75 foram usados no algoritmo.

O Rembrandt é um sistema de reconhecimento de entidades mencionadas (REM), cujo papel no sistema de RIG é a identificação e classificação de todas as entidades mencionadas (EM) presentes na coleção de documentos, e o seu mapeamento a recursos da DBpédia (Cardoso, 2008b; Cardoso, 2012). O Rembrandt guarda os documentos anotados e as EM reconhecidas na base de dados Saskia.

A Saskia organiza as EM em tabelas relacionais para facilitar a geração de índices por parte do indexador (um por cada tipo de EM). A Saskia também serve como API para diversas fontes de informação como a DBpédia, servindo diversos componentes do sistema como é o caso do raciocinador do Renoir. Adicionalmente, a Saskia também associa e armazena informação geográfica da GeoPlanet (Yahoo!, 2011) às EM classificadas como locais, para a posterior geração de índices geográficos, tal como descrito em (Cardoso e Silva, 2010b).

O módulo LGTE (Lucene with GeoTemporal Extensions) (Machado, 2009) é responsável pela recuperação de documentos, e usa o algoritmo BM25 para calcular a similaridade não só entre termos, mas também entre entidades.

4 Participação

A participação no Págico, tal como planeada inicialmente, revelou-se uma tarefa consideravelmente complexa de colocar em prática. A anotação da coleção pelo Rembrandt revelou-se uma tarefa morosa, tal como a rectificação dos padrões de deteção do Renoir para a totalidade dos 150 tópicos.

A participação resumiu-se então à geração de saídas usando o sistema de RIG nas suas configurações mais simples. Estas saídas base serão posteriormente usadas para aferir a diferença de desempenho do sistema, quando a anotação da colecção ficar concluída, e o Renoir conseguir gerar respostas para a maioria dos tópicos.

Assim sendo, as três corridas enviadas são corridas de base:

Saída #1 , sem nenhum tipo de reformulação de consulta, e usando um índice de termos radicalizados da coleção, e com conversão de diacríticos.

Saída #2 , sem nenhum tipo de reformulação de consulta, usando um índice de termos não radicalizados da colecção, e sem conversão de diacríticos.

Saída #3 , com reformulação de consulta, e usando o índice com termos radicalizados e conversão de diacríticos. A reformulação foi feita por com o algoritmo de retorno de relevância cego (*blind relevance feedback*), usando os 10 primeiros documentos no retorno, e adicionando os 16 termos mais relevantes.

O radicalizador usado foi o Snowball (Agichtein e Gravano, 2000). As saídas foram limitadas a 1000 documentos por tópico, e os resultados estão apresentados na tabela 1.

Os resultados mostram que o desempenho do sistema de RIG foi muito fraco, o que reflete o quão inadequado é a utilização de um sistema de RI nesta tarefa sem o auxílio de qualquer tipo de estratégia semântica na procura das respostas.

Mesmo com valores baixos, pode-se observar que as saídas com os índices de termos lematizados geram melhores resultados, uma vez que os termos são agrupados no seu lema, e a recolha é menos sensível às flexões das palavras.

O algoritmo de retorno de relevância cego gera resultados piores, uma vez que é um algoritmo que depende imenso da qualidade dos documentos retornados; se estes são, na sua grande maioria, documentos irrelevantes para o tópico, então os termos gerados poderão dispersar ainda mais o foco da recolha.

5 Ilações da participação

A primeira ilação a retirar desta participação, é a nível técnico: o sistema de RIG usado tem problemas de escalabilidade na anotação. A anotação dos documentos permite gerar um dicionário de EM na coleção e permite usar EMs e as suas classificações semânticas no processo de selecção e recolha documentos. Contudo, a coleção do Págico é composta por mais de 500.000 documentos de tamanho considerável, pelo que a anotação requer grandes recursos computacionais. O Rembrandt é um sistema focado na qualidade de anotação, não na rapidez.

A segunda ilação prende-se com a fragilidade da estratégia do Renoir, nomeadamente da sua dependência num conjunto de manuais na interpretação e raciocínio de consultas. O sistema de RAP desenvolvido pela Priberam também adota uma estratégia semelhante, e que também requer um esforço considerável para a afinação dos seus padrões sintáticos, de forma a conseguir as elevadas prestações obtidas nas avaliações conjuntas em que participou (Amaral et al., 2009).

Além da fragilidade na interpretação das perguntas, geração dos objetos e raciocínio das respostas, a DBpédia ainda pode ser considerada um projecto em franco desenvolvimento. Cada versão da DBpédia é gerada com base em diferentes instantâneos da Wikipédia inglesa; apesar de recentemente a DBpédia disponibilizar bases de dados com base nos instantâneos da Wikipédia portuguesa, o módulo Saskia ainda não está preparado para o utilizar.

Adicionalmente, a DBpédia introduz também alterações significativas nas suas bases de dados, como por exemplo a criação de novas propriedades, ou a revisão das suas classes ontológicas. Tais revisões tornam as regras de raciocínio do Renoir rapidamente obsoletas.

Estas dificuldades técnicas são incontornáveis em qualquer sistema que se proponha desempenhar a tarefa do Págico. Tais sistemas precisam de: i) obter diversas informações sobre uma grande quantidade de entidades em frações de segundos, usando um leque de recursos como almanaques, ontologias ou bases de dados de conhecimento como a DBpédia; ii) compreender as perguntas apresentadas em linguagem natural, e usar a informação disponível para raciocinar e obter respostas, tal como um humano. Veja-se o exemplo do sistema de RAP Watson da IBM (Ferrucci, 2011).

Assim sendo, e uma vez que as dificuldades técnicas estarão sempre presentes, resta mitigar esses problemas e focar na avaliação da estratégia semântica, e na sua aplicabilidade num sistema de RIG. Como trabalho futuro, o Rembrandt será adaptado para anotações mais rápidas e menos completas, para poder lidar com coleções de grande tamanho. Para concluir a avaliação no Págico, um subconjunto de 15 tópicos do Págico (mais relevantes para os objectivos de avaliação de RIG) serão usados para a avaliação pós-Págico.

Agradecimentos

Agradeço a Diana Santos e a José João de Almeida, pelos comentários e sugestões de melhoria deste artigo. Este trabalho foi suportado pela FCT pelo financiamento anual ao LASIGE, projecto GREASE-II (PTDC/EIA/73614/2006) e bolsa de doutoramento SFRH/BD/45480/2008, governo português, União Europeia (FEDER e FSE) através do projecto Linguateca, segundo o contracto ref. POSC/339/1.3/C/NAC, UMIC e FCCN.

Referências

Agichtein, Eugene e Luis Gravano. 2000. Snowball: Extracting Relations from Large Plain-Text Collections. Em *Proceedings of the 5th ACM Conference on Digital Libraries (DL'00)*, pp. 85–94, San Antonio, TX, EUA, June 2-7, 2000. ACM.

Amaral, Carlos, Adán Cassan, Helena Figueira,

André Martins, Afonso Mendes, Pedro Mendes, José Pina, e Cláudia Pinto. 2009. Priberam's question answering system in qa@clef 2008. Em *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, pp. 337–344, Berlin, Heidelberg. Springer-Verlag.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, e Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. (4825):722–735.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de doutoramento, University of Aarhus, Aarhus, Dinamarca, November, 2000.

Cardoso, Nuno. 2008a. Novos rumos para a recuperação de informação geográfica em português. Em Diana Santos, editor, *Linguateca: 10 anos. Encontro satélite do PROPOR 2008*, Aveiro, Portugal, 11 de Setembro, 2008. Linguateca.

Cardoso, Nuno. 2008b. REMBRANDT - Reconhecimento de entidades mencionadas Baseado em relações e análise detalhada do texto. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Cardoso, Nuno. 2012. Rembrandt - a named-entity recognition framework. Em *Proceedingds of the eighth International Conference on Language Resources and Evaluation, LREC 2012*, Istambul, Turquia, 21–27 de Maio, 2012. a aguardar publicação.

Cardoso, Nuno, David Cruz, Marcirio Chaves, e Mário J. Silva. 2008. Using Geographic Signatures as Query and Document Scopes in Geographic IR. 5152:802–810.

Cardoso, Nuno e Diana Santos. 2008. To separate or not to separate: reflections about GIR practice. Em *Proceedings of the 1st Workshop on Novel Methodologies for Evaluation in Information Retrieval, NM EIR'2008*, Glasgow, UK, 30 March, 2008.

Cardoso, Nuno e Mário J. Silva. 2010a. Experiments with Semantic-flavored Query Reformulation of Geo-Temporal Queries. Em *Working Notes of the 8th NTCIR Workshop*, Tóquio, Japão, 15–18 de Junho, 2010.

Cardoso, Nuno e Mário J. Silva. 2010b. A GIR Architecture with Semantic-flavored Query Reformulation. Em *6th Workshop of Geographic Information Retrieval, GIR 10*, Zurique, Suíça, 18-19 de Fevereiro, 2010.

Ferrucci, David A. 2011. Ibm's watson/deepqa. *SIGARCH Computer Architecture News*, 39(3).

Gey, Fredric, Ray Larson, Mark Sanderson, Kerstin Bishoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Di Nunzio, e Nicola Ferro. 2007. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. 4730:852–876.

Machado, Jorge. 2009. LGTE: Lucene Extensions for Geo-Temporal Information Retrieval. Em *Workshop on Geographic Information on the Internet Workshop (GIIW), held at ECIR 2009*, Toulouse, França, 9 de Abril, 2009.

Peters, Carol e Martin Braschler. 2001. Cross-Language System Evaluation: the CLEF campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067–1072.

Rachel Aires, Sandra Aluísio, Paulo Quaresma, Diana Santos, e Mário J. Silva. 2003. An initial proposal for cooperative evaluation on information retrieval in Portuguese. Em Jorge Baptista, Isabel Trancoso, Maria das Graças Volpe Nunes, e Nuno J. Mamede, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003*, pp. 227–234, Faro, Portugal, Junho, 2003. Springer Verlag.

Robertson, Stephen E, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, e Marianna Lau. 1992. Okapi at TREC-3. Em *Proceedings of the 3rd Text REtrieval Conference*, pp. 21–30, Gaithersburg, MD, USA.

Santos, Diana, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Julia Schulz, Yvonne Skalban, e Erik Tjong Kim Sang. 2010. GikiCLEF: Crosscultural Issues in Multilingual Information Access. Em Nicoletta et al. Calzolari, editor, *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, Maio, 2010. European Language Resources Association (ELRA).

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.

Santos, Diana, Nuno Cardoso, e Luís Miguel Cabral. 2010. How geographic was gikiclef?: a gir-critical review. Em Ross Purves, Paul Clough, e Christopher B. Jones, editores, *GIR*. ACM.

Santos, Diana, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, e Yvonne Skalban. 2009. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer.

Santos, Diana, Paula Carvalho, Hugo Oliveira, e Cláudia Freitas. 2008. Second HAREM: new challenges and old wisdom. (5190):212–215.

Santos, Diana, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela, e Susana Afonso. 2004. Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa. Em Guillermo De Ita Luna, Olac Fuentes Chávez, e Mauricio Osorio Galindo, editores, *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanol y el Portugués" and IX Iberoamerican Conference on Artificial Intelligence, IBERAMIA 2004*, pp. 147–154, Puebla, México, Novembro, 2004.

Silva, Mário J., Bruno Martins, Marcirio Chaves, Ana Paula Afonso, e Nuno Cardoso. 2006. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban Systems*, 30:378–399.

Voorhees, Ellen M. e Donna Harman, editores. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

Yahoo! 2011. GeoPlanet^{TM}Key Concepts. `http://developer.yahoo.com/geo/geoplanet/guide/concepts.html`. acedido em Fevereiro de 2012.

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/`.

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.

Desafios na recolha de informação baseada na Wikipédia portuguesa com o Págico

João Miranda
Instituto Superior Técnico
joaocarvalhomiranda@ist.utl.pt

Resumo

O Págico foi uma iniciativa de recolha de informação em português, em que se usou uma cópia local da Wikipédia portuguesa para responder a 150 tópicos sobre temas referentes à lusofonia. As perguntas não tinham um número limitado de respostas previamente conhecido. O sistema de apoio ao Págico permitia navegar e pesquisar na cópia local da Wikipédia e apresentar as respostas e justificações aos tópicos. Este artigo sumariza os principais desafios encontrados e a metodologia usada com a participação humana nesta iniciativa.

Palavras chave

Avaliação conjunta, recolha de informação, Págico, lusofonia, Wikipédia

1 Introdução

O Págico foi organizado pela Linguateca e surgiu como sequência do GikiCLEF (Santos et al., 2010), uma iniciativa de recolha de informação em diferentes línguas. Ao contrário do GikiCLEF, o Págico foca-se apenas numa língua, o português, utilizando uma versão da Wikipédia portuguesa e perguntas em português a temas de cariz lusófono.

Um dos objectivos do Págico era avaliar e poder comparar o desempenho dos sistemas de resposta automática, e também humana, a perguntas para as quais não há um número limitado de respostas previamente conhecido. As respostas tinham de ser justificadas, o que significa que não bastava indicar as respostas, era necessário justificá-las. As respostas a dar eram os próprios artigos da Wikipédia correspondentes à resposta pretendida: a resposta tinha de ser ela própria uma entrada da Wikipédia e não páginas onde a resposta estivesse presente. Por exemplo, algumas das aves descritas na página *Aves de Angola* da cópia local da Wikipédia não tinham página própria criada e não podiam, por isso, ser dadas como resposta ao tópico *Aves de Angola*.

A colecção de avaliação disponibilizada pelo Págico tinha 150 tópicos para resposta. A cada tópico correspondia um, ou mais, dos seguintes temas: Artes, Ciência, Cultura, Desporto/esportes, Economia, Geografia, Letras, Política. Cada tópico tinha atribuído um ou mais dados geográficos que variavam entre: Angola, Brasil, CaboVerde, Geral, GuinéBissau, Lusofonia, Macau, Moçambique, Portugal, SãoToméPríncipe, Timor.

O Págico estava assente no sistema SIGA (Costa, Mota e Santos, 2012) que permitia visualizar os tópicos, navegar e pesquisar na versão local da Wikipédia, e apresentar as respostas aos tópicos e justificações seleccionadas. Foi utilizada uma cópia da Wikipédia portuguesa do dia 25 de Abril de 2011.

Na secção 2 deste artigo mencionam-se alguns trabalhos anteriores a esta iniciativa. Na secção 3 apresenta-se a motivação para a participação humana no Págico. Na secção 4 descreve-se a metodologia utilizada para responder aos tópicos. Na secção 5 referem-se os desafios encontrados. Na secção 6 apresentam-se algumas conclusões e breves sugestões de melhoria do sistema.

2 Trabalho relacionado

A Linguateca organizou anteriormente várias iniciativas de avaliação conjunta. A iniciativa predecessora do Págico, o GikiCLEF, foi organizada em 2009 no âmbito do CLEF[1]. O GikiCLEF disponibilizava 50 tópicos em 9 línguas europeias, para os quais não havia um número determinado de respostas conhecidas. O GikiCLEF surgiu na sequência do GikiP (Santos et al., 2009), organizado um ano antes e que continha apenas 15 tópicos em 3 línguas europeias.

[1] http://www.clef-initiative.eu

3 Motivação

A motivação para a participação humana no Págico compreendeu diferentes pontos. Por um lado, poder participar num desafio de avaliação conjunta é, por si só, interessante. Por outro lado, é estimulante pôr à prova as capacidades de recolha de informação num sistema de pesquisa limitado. Aprender coisas novas em matérias fracamente dominadas é, também, enriquecedor.

Outra das motivações foi responder ao desafio da luta homem-máquina que sempre despertou o interesse da Inteligência Artifical.

4 Metodologia

Na busca de informação as técnicas de pesquisa podem basear-se nos termos de partida ou nos termos de chegada. A diferença é subtil mas muito relevante e baseia-se na distinção entre aquilo por que queremos procurar e aquilo que queremos encontrar. Por exemplo, pesquisar por *aves de Angola* é diferente de pesquisar por *"é uma ave de Angola"*: a segunda opção permite obter resultados directos, se os houver.

Há quatro cenários principais na busca de informação que influenciam o sucesso de uma pesquisa:

1. sabemos onde está determinada informação e sabemos identificá-la;
2. sabemos onde está determinada informação mas não sabemos identificá-la;
3. não sabemos onde está determinada informação mas sabemos identificá-la;
4. não sabemos onde está determinada informação nem sabemos identificá-la.

É particularmente difícil encontrar informação quando não sabemos onde ela está nem sabemos identificá-la. É o caso de quando não se conhece a resposta a um tópico nem se vislumbram quais os artigos que nos poderão ajudar a respondê-lo. Por motivos de gestão do tempo de resposta, os tópicos que correspondiam a este cenário foram relegados para análise posterior, que não chegou a acontecer, em favor dos que se enquadravam nos três primeiros cenários.

Dos 40 tópicos respondidos, 27,5% enquadravam-se no primeiro cenário, 47,5% no segundo e 25% no terceiro. Por exemplo, o tópico *Línguas faladas em Timor Leste* enquadrava-se no primeiro cenário: conheciam-se duas respostas possíveis de antemão e bastava apenas verificar se existiam os artigos correspondentes na cópia local da Wikipédia.

Não houve uma ordenação intencional definida na escolha dos tópicos a responder. Na progressão das respostas às perguntas, para além de se preterirem as que correspondiam ao cenário 4, seguiram-se, em geral, as seguintes linhas de orientação:

1. se o tema de uma dada pergunta era familiar partia-se para a resposta pesquisando pelo artigo tido como provável de conter a resposta;
2. se o tema não era familiar, tentava-se uma pesquisa com uma expressão de busca contendo um dos termos ou expressões existentes na pergunta;
3. se uma pesquisa não oferecia, à segunda tentativa, resultados satisfatórios, partia-se para uma nova pergunta.

Como as perguntas eram de complexidade diferente, enquanto para algumas a resposta foi obtida navegando por poucos artigos, houve outras em que foi necessário consultar mais artigos para chegar às respostas e às justificações. Por exemplo, o tópico *Países que venceram a Copa do Mundo em uma disputa de pênaltis* foi respondido com 2 respostas e 2 justificações diferentes, enquanto o tópico *Escritores cabo-verdianos com obra publicada em crioulo* foi respondido com 4 respostas e nenhuma justificação. O grau de complexidade das perguntas era influenciado por diferentes factores, como o nível de familiaridade com o tema, o número de respostas a dar, o número de páginas que era necessário consultar e cruzar para responder a um tópico e justificar a resposta, ou o tempo dispendido até encontrar uma resposta considerada correcta. Dos 40 tópicos respondidos, 22,5% foram considerados fáceis, 27,5% de dificuldade média e 50% foram considerados difíceis.

Depois de respondidas, as respostas e justificações foram revistas para confirmação de correcção.

Para a pesquisa não foi dada particular atenção ao *Tema* e *Dados geográficos* de cada tópico. Foram usados, essencialmente, os termos, expressões e entidades mencionadas presentes em cada tópico.

5 Desafios encontrados

Houve diferentes desafios encontrados com a participação no Págico. Em primeiro lugar, foi necessária a familiarização com os termos habitualmente utilizados nesta área (ex.: *tópicos* e *corridas*).

Em segundo lugar, houve uma dificuldade inicial transitória em perceber o que devia ser apresentado como resposta e o que devia ser apresentado como justificação. O triângulo pergunta-resposta-justificação seguia um formato próprio e as respostas não eram dadas como num jogo de perguntas vulgar. As respostas eram os próprios artigos da cópia local da Wikipédia, e não artigos que pudessem conter a resposta, ou que a permitissem deduzir de forma indirecta. Nem todos os tópicos correspondiam a perguntas feitas de forma interrogativa. Enquanto umas eram interrogações directas (ex.: *Quem descobriu São Tomé e Príncipe?*) outras eram feitas de forma indirecta (ex.: *Empresários lusófonos com uma fortuna considerável*). Na Tabela 1 apresentam-se alguns exemplos de tópicos do Págico, e respectivos *Temas* e *Dados geográficos*.

As respostas não eram, em geral, directas, isto é, os tópicos foram construídos de forma a que fosse necessário relacionar artigos para encontrar uma resposta e poder justificá-la usando outros artigos, consoante as necessidades. A ideia era retirar o ênfase da extracção de respostas directamente a partir do texto e testar a capacidade de resposta quando é preciso cruzar informação de artigos diferentes.

As limitações do sistema de pesquisa foram uma das maiores dificuldades encontradas. O sistema de pesquisa apenas permitia procurar no título da página e pela ordem dos termos introduzidos na expressão de busca. Para responder às perguntas, ou se sabia de antemão a resposta à pergunta e se procurava o respectivo artigo, ou se iniciava a pesquisa com uma expressão de busca que se julgasse ser um bom ponto de partida para encontrar uma resposta. Como as respostas não eram fechadas, ou seja, não havia um número prévio limitado de respostas conhecidas, era por vezes difícil decidir se uma pergunta estava satisfatoriamente respondida e justificada; e se se deveria partir para outra pergunta ou, antes, procurar mais respostas e justificações para a pergunta corrente. Por exemplo, ao tópico *Instrumentos musicais de origem africana comuns no Brasil* deram-se 1 resposta e 1 justificação, enquanto ao tópico *Telenovelas brasileiras passadas no tempo da escravatura no Brasil* se deram 4 respostas e nenhuma justificação. Por seu turno, ao tópico *Aves de Angola* apresentaram-se 6 respostas e 1 justificação igual para todas as respostas. Seriam 6 respostas suficientes? Bastariam 3, ou 10 seria melhor?

A Wikipédia original pode ser usada como recurso de tradução: sabendo um termo numa língua de partida, é possível usar a correspondência de artigos entre línguas para encontrar o termo tido como equivalente numa língua de destino. O mesmo se passa ao procurar informação cruzada entre artigos de diferentes línguas, uma vez que os artigos têm, frequentemente, informação e completude variada entre elas. Por isso, ao pesquisar na Wikipédia é, muitas vezes, vantajoso partir de uma língua diferente para encontrar a informação pretendida. O cruzamento de artigos em mais do que uma língua permite uma abrangência maior de informação que falta na versão monolingue usada no Págico.

A Wikipédia oferece outras capacidades de cruzamento e extracção de informação: desde as hiperligações entre artigos até à categorização e hierarquia de artigos. As categorias são uma funcionalidade que permitiria resposta facilitada a muitos tópicos, mas tendo sido esvaziadas na versão utilizada no Págico, revelaram-se inúteis: as categorias existiam mas não continham informação. Não era, por isso, possível ir à categoria *Aves de Angola* para responder ao tópico *Aves de Angola*. Tornava-se, pois, mais difícil relacionar informação, o que seria feito com relativa facilidade se se pesquisasse a mesma informação na Wikipédia original.

De um ponto de vista mais geral, em qualquer matéria de estudo as tarefas são facilitadas se houver um fio condutor: desde a investigação criminal, aos processos de memorização cerebral, à comunicação entre diferentes unidades de uma empresa. Partir esta interligação de informação na versão local da Wikipédia é reduzir as capacidades de sucesso de resposta aos tópicos, uma vez que obriga a navegar e a ler os artigos de forma mais exaustiva. Significa, também, colocar carga adicional de processamento e tempo dispendido na pesquisa de páginas relacionadas com o tema que se procura. Do ponto de vista humano, este tempo pode não ser aceitável quando se pretende responder a uma pergunta de forma completa. Na verdade, se o objectivo era obrigar a avaliação humana a cingir-se à pesquisa de páginas isoladamente e a relacionar depois a informação entre si, o objectivo foi, dessa forma, atingido. Mas é um retrocesso na forma como nos habituámos a lidar com a informação e a organizá-la para tornar mais simples a pesquisa e navegação. O ser humano tem capacidades notáveis de relacionamento e cruzamento de informação mas tem limitações no que respeita à quantidade de dados que pode analisar ao mesmo tempo. Para um humano, é difícil gerir muita informação em simultâneo. E esse é o impacto mais visível que os computadores introduziram em diversos campos,

	Tópicos	Temas	Dados geográficos
Pagico_008	Telenovelas brasileiras passadas no tempo da escravatura no Brasil	Artes, Letras	Brasil
Pagico_027	Doenças letais comuns em países lusófonos transmitidas por mosquitos	Ciência	Lusofonia
Pagico_098	Cidades dos Estados Unidos que tiveram forte imigração portuguesa	Geografia, Letras	Portugal
Pagico_119	Pratos típicos da gastronomia de Cabo Verde	Cultura	Cabo Verde
Pagico_135	Aves de Angola	Ciência, Geografia	Angola

Tabela 1: Exemplo de tópicos no Págico

incluindo o da Linguística Computacional.

Entre os pontos favoráveis do processo de resposta incluíam-se a ausência de tempo limite para responder a uma pergunta e a não contabilização desse tempo para efeitos de avaliação de desempenho.

Por último, o elevado número de perguntas fez antever que seria difícil responder a todos os tópicos em pouco tempo. Este foi o principal factor que levou a que não tenham sido todos respondidos. Preferiu-se, também, procurar responder a um número mais reduzido de tópicos mas com uma maior completude de respostas e justificações julgadas correctas.

6 Conclusões

As ideias principais a reter com a participação nesta iniciativa estão relacionadas com as principais dificuldades sentidas:

1. as categorias tinham sido esvaziadas, o que inibia a resposta imediata a perguntas em que a informação podia ser extraída directamente das páginas de categoria;

2. mais importante, isso dificultava a navegação contínua entre páginas com pontos comuns entre si, uma vez que as categorias também servem para facilitar a agregação de páginas relacionadas, e isso é uma boa ajuda para extrair informação (que não constando desta versão modificada da Wikipédia, consta da versão original).

Humanamente, é difícil adivinhar que páginas conterão determinada resposta nos casos em que não se domina a matéria analisada, e isso notou-se muito ao procurar a informação usando a interface de pesquisa disponibilizada no Págico. Essa função é facilitada pelos motores de busca, e é a diferença óbvia entre:

1. pesquisar qualquer informação numa enciclopédia em papel, onde é preciso saber em que entradas se irá procurar o que pretendemos, ou seja, é preciso saber por onde começar, ou:

2. pesquisar num motor de busca avançado, como o Google, onde esse requisito não é importante.

Na primeira é necessário prever as entradas, e parte-se das entradas para os conteúdos; na segunda, faz-se o caminho inverso e descobrem-se facilmente as entradas a partir dos conteúdos.

Sentiu-se que seria mais fácil usar um motor de busca avançado para descobrir que páginas da Wikipédia conteriam determinada informação do que vaguear ao acaso entre artigos na esperança de encontrar as páginas que respondessem ao que se precisava. Da mesma forma que numa enciclopédia em papel seria necessário percorrer os artigos que se considerasse levarem à informação pretendida, também aqui na versão da Wikipédia do Págico foi sentida essa imposição. Mesmo que a informação estivesse lá, poderia não ser descoberta por não se saber onde procurá-la.

Do ponto de vista da luta homem-máquina, é aí que um sistema avançado de recuperação de informação ganha a um humano: o sistema encontra a informação em segundos e é capaz de procurar em toda a enciclopédia num instante, e um humano não é capaz de o fazer. Enquanto um humano anda de artigo em artigo a descobrir que missionários estiveram no Brasil no tempo dos Descobrimentos, um bom motor de busca faz isso num instante: tem as páginas que falam de missionários todas indexadas, todas em "memória", e o humano não.

Mas há uma coisa em que os humanos são melhores do que um motor de busca avançado: podem jogar com as expressões de busca e conduzir a pesquisa como querem, e um motor de busca não saberia nunca fazer isso. Em última instância são os humanos que verificam se os resultados que ele dá correspondem ao que procuram e contêm a resposta pretendida. São os humanos que refinam as pesquisas se entenderem que os resultados

não correspondem às expectativas. Um motor de busca avançado sabe que os resultados que dá têm aquilo por que se procurou, mas não sabe se o utilizador humano encontrou aquilo que de facto procurava.

No geral, o que se sentiu foi uma regressão nas capacidades de pesquisa. Encontrou-se extrema dificuldade em encontrar informação que se encontraria de forma fácil com outros instrumentos. Talvez seja reflexo da habituação às facilidades que os motores de busca vieram trazer.

Há três pontos-chave que permitiriam responder com maior sucesso e em menos tempo aos tópicos do Págico:

1. as categorias da Wikipédia;
2. os conteúdos noutras línguas;
3. um sistema de pesquisa que permitisse pesquisar em todo o artigo e não apenas no título.

Estes três pontos reflectem a diferença entre conseguir responder aos tópicos usando apenas o sistema disponibilizado, um sistema controlado do ponto de vista de validação, ou usando as ferramentas livremente disponibilizadas na Internet, mas que não permitiriam igualdade de recursos utilizados do ponto de vista de avaliação conjunta.

A interface disponibilizada pelo SIGA, apesar de simples, é suficiente para aquilo a que se propõe.

Agradecimentos

Agradeço à equipa do Págico a oportunidade de ter participado nesta iniciativa e, em especial, à Cristina Mota pelo acompanhamento e ajuda ao longo do tempo.

Agradeço a apreciação e pertinência das observações de Cláudia Freitas e Stella Tagnin que ajudaram a tornar este artigo mais esclarecedor e interessante.

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.

Santos, Diana, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Julia Schulz, Yvonne Skalban, e Erik Tjong Kim Sang. 2010. Gikiclef: Crosscultural issues in multilingual information access. Em Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may, 2010. European Language Resources Association (ELRA).

Santos, Diana, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, e Yvonne Skalban. 2009. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer, pp. 894–905.

O que é uma resposta? Notas de uns avaliadores estafados

Cláudia Freitas
Linguateca/FCCN & PUC-Rio
maclaudia.freitas@gmail.com

Paulo Rocha
Linguateca/FCCN
paulo.rocha@xadrez64.com

Cristina Mota
Linguateca/FCCN
cmota@ist.utl.pt

Luís Costa
Linguateca/FCCN
luis.f.kosta@gmail.com

Diana Santos
Linguateca/FCCN &
Universidade de Oslo
d.s.m.santos@ilos.uio.no

Resumo

Após argumentar que a avaliação de respostas é algo extremamente complexo, este artigo descreve o processo de avaliação das respostas do Págico, e as questões surgidas no próprio processo, assim como sugere um modelo de gradação entre respostas e a sua possível futura classificação nesses moldes. Além de descrever o processo seguido, o artigo sugere formas mais avançadas de interface de avaliação (a desenvolver no futuro). O problema das justificações e como é preciso melhorar essa questão é também apresentado e discutido. Uma análise dos comentários dos avaliadores, assim como alguns dados quantitativos sobre esses comentários, é depois apresentada.

Palavras chave

Resposta automática a perguntas, avaliação, wikipédia

1 Apresentação

A parte do trabalho de organizar uma avaliação conjunta que é considerada, por parte do público em geral, a menos problemática – e exigindo menos preparação teórica e científica – é sem dúvida a avaliação das respostas (dadas por um paricipante a perguntas feitas pela organização). Tal não é contudo correto, e de facto a razão principal deste artigo é o de argumentar mais uma vez, com dados concretos, a favor da complexidade do problema.

É certo que em cada avaliação conjunta os casos mais problemáticos são mencionados nos artigos correspondentes, mas os leitores provavelmente consideram os casos relatados como casos especiais, interessantes e possivelmente dignos de nota, mas anedotais no sentido de não corresponderem a uma situação sistemática, nem precisarem de uma reflexão aprofundada.

No Págico a situação de avaliação, dado aliás que uma das nossas motivações era perceber melhor a própria procura e justificação de informação, levou a que decidíssemos escrever sobre o processo, como forma de iluminação do próprio assunto que estamos a investigar.

Seja como for, apresentamos nesta introdução alguns outros casos da literatura, não só para mostrar que este é um assunto de interesse geral, como para tornar claro que não pretendemos ter sido os primeiros a identificar este problema, que aliás também já tratámos em ocasisões anteriores (Rocha e Santos, 2007; Santos, 2007; Simões, Rocha e Fonseca, 2009).

A primeira questão é o que considerar uma resposta, ou melhor, informação suficiente para poder considerar uma resposta correta. Como pretendemos tornar claro, a noção de resposta (certa, apropriada ou útil) é muito mais fluida do que seria de esperar.

O que pode parecer óbvio antes de olhar para as respostas, deixa de o ser quando consideramos as possibilidades de obter uma resposta certa mas não esperada nem necessariamente útil. Assim Voorhees e Tice (2000) relata o caso da pergunta "onde é o Taj Mahal", corretamente respondido ... com um endereço em Nova Iorque sobre um restaurante indiano do mesmo nome. E Sparck Jones (2003) refere a resposta à pergunta "quem é o autor do Ivanhoe", corretamente respondida por "o autor de Rob Roy", mas que podia não ser útil para quem não soubesse quem tinha escrito ambos os livros. Ou seja, é útil? Depende de facto de qual era a intenção da pergunta, e do conhecimento de quem a fez. Não existe, portanto, uma resposta única correta, mesmo quando a pessoa que pergunta

disso está convencida.[1] O que era o caso nestes dois exemplos, e que é aliás origem de um género de piadas muito comum (respostas certas mas inesperadas).

Por outro lado, a maior parte das perguntas autênticas podem ser decompostas num conjunto de bocados, que no TREC foram chamados pepitas ("nuggets"), e cuja resposta pode ser avaliada pelo menos parcialmente, vendo quantos bocados de uma resposta complexa ou com necessidade de justificação complexa são obtidos pelos sistemas. Por exemplo, para responder a que rios italianos fluem de norte para sul poderia ser preciso estabelecer que o Pó é um rio, que é italiano, e que flui de norte para sul, e cada um dos três bocados poderia ser julgado independentemente.

Se tivermos um tópico como [realizadores que fizeram filmes sobre a independência do Brasil] podemos ter de avaliar respostas sobre realizadores que fizeram filmes sobre a independência dos EUA e respostas (obviamente completamente erradas, mas autênticas, provenientes de sistemas automáticos!) sobre estados com o nome "Independence". Paradoxalmente, é muito mais fácil avaliar (e rejeitar) a segunda resposta do que a primeira. Ou seja, respostas completamente erradas são mais fáceis de avaliar do que respostas em que haja uma sobreposição de conteúdo que leva à necessidade de uma investigação muito mais pormenorizada. Isto é semelhante à questão dos significados das palavras e à sua tradução: quase-sinónimos são muito mais complicados de distinguir e de formalizar do que sentidos muito distintos (Santos, 2012).

Para mais exemplos de respostas complicadas de avaliar, quer no que se refere à justeza da sua justificação, quer à interpretação não intencional, veja-se Rocha e Santos (2007).

2 O conceito de resposta no Págico

No Págico, dadas as perguntas, ou tópicos, descritas em Freitas (2012), pretendemos que os sistemas ou participantes apresentassem como resposta páginas da coleção do Págico (Simões, Costa e Mota, 2012), adicionalmente associadas a mais páginas da mesma coleção – chamadas justificações ou justificativas – se a verificação dessa página como resposta tivesse de passar por mais informação.

Uma resposta no Págico foi portanto formalmente definida como o título de uma página da wikipédia (na coleção) com o tipo semântico apropriado (se perguntamos por pessoas, não servem filmes, se perguntamos por países, não aceitamos bandeiras, veja-se a secção 5 abaixo), e em que a informação dessa página, eventualmente suplementada com a informação de mais um conjunto de páginas apresentadas como justificativa, permitia a uma pessoa confirmar essa informação.

Por não termos suficiente conhecimento do problema e da forma como excesso ou confirmação de justificativas influenciaria uma pessoa genuinamente interessada nas respostas, não nos pronunciámos sobre eventuais penalizações ou prémios por justificações redundantes.

Apenas indicámos claramente que uma resposta certa, mas não justificada, não contaria para o desempenho dos sistemas.

3 O processo de avaliação

Como indicado em Mota (2012), obtivemos 52879 respostas (candidatas a respostas) correspondendo a 32485 respostas diferentes. Apenas os participantes humanos apresentaram respostas com justificação, os sistemas automáticos apenas apresentaram respostas "auto-justificadas", no sentido de que não precisariam de mais informação para serem consideradas corretas.

As respostas foram distribuídas pelos avaliadores (os autores do presente artigo), que fizeram a avaliação caso a caso. Os casos que suscitaram dúvidas foram posteriormente discutidos pela organização. O número de respostas avaliadas por cada autor divergiu muito, tendo cabido a Paulo Rocha a parte de leão.

Embora o SIGA (Santos e Cabral, 2009) permita que uma resposta seja avaliada por vários avaliadores, ajudando depois à resolução de conflitos, algo aliás que o tornou pioneiro na gama dos sistemas de apoio à avaliação[2], não tivemos infelizmente tempo no Págico para fazer isto extensivamente: de facto, apenas as respostas marcadas como duvidosas foram avaliadas por mais de um avaliador, e em metade

[1] O Alberto Simões chamou a atenção para que esta é uma afirmação forte demais, mas notamos que "não existe sempre", ou "não existe algumas vezes" seriam fracas demais, porque dariam a entender que a organização, ou quem pergunta, se deveria esforçar mais, o que achamos que não é o caso. Perguntas autênticas, feitas por pessoas em casos naturais, são sempre vagas e susceptíveis da interpretações não antecipadas, e esta parece-nos uma característica suficientemente interessante da linguagem e da comunicação humanas para não almejarmos uma precisão exagerada.

[2] Como referido em (Santos et al., 2010, página 2350), os outros sistemas esperam apenas uma avaliação por resposta.

dos casos, por uma lapso, o segundo avaliador teve acesso / soube da avaliação do primeiro. Ambas estas questões foram devidas ao muito reduzido prazo, em tempo de quadra natalícia, que tivemos para efetuar a avaliação.

É preciso de qualquer maneira salientar que os avaliadores não eram necessariamente especialistas sobre os variados tópicos, e em muitos casos, por desconhecimento do assunto ou de particulares casos concretos, não lhes era fácil avaliar uma resposta. Nesses casos contudo foram encorajados a deixar um comentário, ou a perguntar diretamente ao criador do tópico questões de clarificação.

As respostas duvidosas podem ser distribuídas por uma variedade de "classificações", a saber

- a resposta parecia certa ao avaliador mas não havia justificação – e nem sempre um avaliador é tão conhecedor de um assunto que pode confiar totalmente na sua erudição. Se instado a provar que é certa e não apenas que acha que é certa, provavelmente teria de ir fazer investigação sobre o assunto, o que nos quadros dos prazos de avaliação do Págico estaria completamente fora de questão

- a justificação não era muito aceitável – ou seja, não convenceu completamente o avaliador, mas isso podia ser devido a diferente conhecimento sobre o assunto, ou mesmo diferente opinião sobre o assunto. Por exemplo no **tópico 44** [Lendas ou personagens folclóricas de origem indígena] havia casos em que estava indicado que não se conhecia a origem da lenda, ou que havia explicações alternativas.

- a classificação da página era um pouco ao lado, o que significa que a resposta podia estar contida na página, mas a página era sobre outra coisa

- partes da justificação eram apenas subentendidas, ou exigiam conhecimento complicado – por exemplo, é suficiente ler que estamos em presença de uma cidade raiana? "Raiana" significa, em Portugal, "perto da fronteira com a Espanha", mas é pouco provável que noutros países lusófonos essa denominação seja conhecida

- a justificação ou parte dela estaria em figuras ou tabelas (infoboxes) que não se encontravam na coleção do Págico[3]

Ainda existe, contudo, um acervo grande de comentários que podem ser explorados e garimpados para maior compreensão dos problemas, e que estamos a considerar talvez vir a tornar público após uma revisão e sistematização dos mesmos.

4 Tópicos ambíguos ou vagos e as consequências nas respostas

Em alguns tópicos, percebemos já antes do processo de avaliação a ambiguidade, ou vagueza, do que perguntamos. Por exemplo, veja-se o **tópico 61** [Movimentos culturais em países lusófonos que se refletiram nas artes plásticas e na música], o **tópico 75** [Organizações ou grupos armados que lutaram contra o regime militar no Brasil] e o **tópico 64** [Igrejas do Rio de Janeiro construídas por irmandades ou confrarias de negros].

No primeiro exemplo temos dois pontos pouco claros: como não explicitamos que os movimentos deveriam ser originários de países lusófonos, aceitamos qualquer movimento que se refletisse nas artes plásticas e na música. Além disso, como também não explicitamos que nos interessava a interseção – tanto nas artes plásticas como na música – aceitamos a disjunção.

No segundo exemplo, embora a intenção fosse encontrar organizações e grupos, ambos armados, aceitamos igualmente a leitura em que "armados" refere-se apenas aos grupos.

Por fim, no tópico 54, como não especificamos se o interesse estava no estado ou na cidade do Rio de Janeiro, decidimos aceitar respostas com ambas interpretações.

Note-se que estas são dúvidas gerais que se puseram aos avaliadores, não necessariamente com base em respostas concretas.

E o que dizem os resultados nesses tópicos? Essas são questões relevantes?

- Em relação a ser armado ou não, nas 16 respostas que considerámos corretas, houve quatro casos não armados: dois partidos, um que refere explixcitamente "naõ armado"; e outro que não se refere a armas. Donde se conclui que esta especificação é importante

[3] Estritamente falando, do ponto de vista da avaliação da própria wikipédia, poderia fazer sentido separar a situação de estar no instantâneo usado, ou estar na versão atual, como notado pelo Alberto Simões, mas na prática ninguém usou o instantâneo de 25 de abril: ou usaram a coleção feita por nós, ou a wikipédia corrente à data do Págico.

e, se fosse fulcral para o participante, devia ter sido mais rigorosamente exprimida, resultando assim em apenas 12 e não 16 respostas.

- Em relação aos movimentos culturais, das 256 respostas obtidas, houve 12 que foram consideradas corretas. Dessas não conseguimos encontrar nenhuma que fosse apenas musical, mas dez mencionam explicitamente a música também, sendo que três delas tem expressão primordial (ou origem) na música. Constatamos portanto que havia pouca diferença entre exigir tanto na música como nas artes plásticas, sendo que o número de respostas justificadas passaria de 12 para 10 apenas.

- Em relação à questão da localização estado ou cidade, que aliás e uma fonte de problemas para sistemas de recolha de informação geográfica (RIG), visto que as capitais de um estado têm o mesmo nome que o dito, não pudemos tirar qualquer conclusão, pois das 219 propostas pelos participantes houve apenas uma resposta correta, em que há referência explícita à localização na cidade do Rio de Janeiro.

Um caso que consideramos interessante diz respeito ao **tópico 18** [Discos brasileiros considerados marcantes na história da música brasileira]. Embora a formulação seja clara, sabemos que "marcante" é uma característica altamente subjetiva, o que não impede que esta seja uma pergunta autêntica no sentido de ser comum, e nos interessa perceber como o adjetivo foi "traduzido" pelos sistemas. Voltaremos a tratar desse exemplo na secção 7.

Finalmente, algo que detetámos durante a avaliação foi a questão do uso, provavelmente exagerado, dos termos "lusofonia" ou "lusófonos" na formulação dos tópicos, que levou por vezes a participação automática a produzir resultados completamente espúrios. Talvez na nossa avidez de produzir perguntas associadas à lusofonia como um todo tenhamos acabado por criar tópicos artificiais, que não tivessem nenhuma aplicação prática. Com efeito, é pouco provável que o **tópico 147** [Museus em capitais luofonas] fizesse sentido a um usuário normal. Pelo contrário, reconhecemos que "Museus em Lisboa", ou "Museus em Brasília", seriam necessidades de informação muito mais naturais.

5 A questão do tipo da resposta

Uma questão que mantivemos do GikiCLEF (Santos e Cabral, 2009) mas que é seguramente controversa é a exigência de que uma resposta correta tem de ser do tipo subjacente à pergunta.

Por exemplo, numa pergunta como "que países têm amarelo na bandeira", em que uma resposta certa seria "Brasil", sistemas que enviassem a resposta "Bandeira do Brasil" não recebiam qualquer pontuação – ou melhor, essa resposta era implacavelmente considerada errada.

Muitos participantes, contudo, estavam radicalmente em desacordo, argumentando que qualquer pessoa ficaria satisfeita com essa resposta, de facto mais satisfeita do que com a resposta "Brasil", em que teriam de ir à procura da bandeira.

A questão é a seguinte: Embora de um ponto de vista lógico, a resposta estivesse incorreta, de um ponto de vista prático, era até uma resposta melhor. Quando os critérios da lógica e da utilidade não são convergentes, temos de decidir se exigimos ambos, ou se aceitamos apenas um:

- se deixamos apenas a utilidade, aceitando por isso páginas como bandeira do Brasil como resposta, onde paramos, até chegar à tarefa de recolha de informação (RI) simples?

- se deixamos apenas a lógica, podemos ter respostas logicamente corretas mas inúteis, como por exemplo "países com bandeira azul e amarela", como resposta a "Que países têm amarelo na bandeira?".

Por outro lado, temos também de indicar que a especificação do tipo de resposta torna a avaliação (no sentido de recusar respostas de tipo errado) muito mais fácil e rápida, o que é um critério não só importante para os avaliadores mas para os próprios utilizadores, que reconhecem a resposta no título da página da wikipédia em vez de terem de procurá-la nas páginas.

6 O que é uma justificativa?

Sem dúvida, um dos pontos mais controversos da avaliação, e que assumimos ter sido subestimado pela organização, diz respeito ao que é, exatamente, uma justificativa, visto que a noção envolve a mensuração de informações de difícil quantificação, como o quanto de conhecimento

do assunto o formulador da pergunta tem e o quanto de conhecimento partilhado há entre quem formula a pergunta e quem responde.

Deveríamos ter um conjunto de hipóteses que assumimos que todos sabem e não as pedir no caso da participação humana? (diferentemente da participação automática, em que os sistemas teriam de explicitamente apresentar uma justificativa). E que hipóteses seriam essas? Como definir "um conhecimento que todos têm"?

Por exemplo, se um tópico envolve "capitais de países lusófonos", deveríamos exigir que o participante acrescentasse, como justificativa, uma página com a informação de que Brasília é a capital do Brasil (se isso não estivesse já mencionado na página de resposta, naturalmente)? E, ainda, uma página com a informação de que no Brasil se fala português? Ou podemos considerar todas essas informações já assumidas, e portanto bastaria a menção, no texto, a alguma capital de país lusófono? Onde deveríamos parar com a exigência da justificação?

Para um tópico como [Movimentos culturais surgidos no nordeste do Brasil], é fácil imaginar que um participante – pessoa – brasileiro consideraria como resposta autojustificada uma página que localiza o movimento em Recife, visto ser perfeitamente óbvio que Recife fica no nordeste. Para aqueles que não têm ideia da localização de Recife, a informação da localização de Recife é relevante, e portanto a resposta precisaria de justificativa.

O mesmo se aplica a [Séries ou minisséries brasileiras baseadas em romances portugueses]. Se a página informa que a série é baseada no romance Os Maias, de Eça de Queiroz: é preciso a informação explícita de que Eça de Queiroz é um autor português? Certamente a noção de justificativa esbarra no conhecimento prévio dos participantes.

No caso da participação dos sistemas, exigimos sempre justificação, pois consideramos que não podemos confiar num conhecimento prévio de sistemas, ou que de qualquer maneira sistemas automáticos não são capazes de decidir o que é óbvio ou não, e terão de deixar essa decisão aos seus utilizadores. Mas assumimos aqui que talvez estejamos misturando duas noções: "justificativa" e "necessidade de confirmar o raciocínio automático".

De fato, a questão "o que é uma justificativa" não é consensual nem mesmo na organização, e apresentamos aqui alguns pontos que, acreditamos, são merecedores de discussão mais aprofundada:

1. justificativas que parecem desnecessárias a seres humanos falantes de português, visto que apenas parafraseiam a pergunta, ou que pressupõem esse conhecimento para serem respondidas (por uma pessoa). Ou seja, dados do tipo "Se Luanda, então Angola"; "Se Eça de Queiroz, então português". São situações que envolvem um conhecimento estável, como a relação entre países e suas capitais, e a nacionalidade de pessoas, entre outras. Expresso de outra forma, casos em que ficaríamos contentes com uma resposta que não explicitasse isso.
2. justificativas que não podem ser (logicamente?) inferidas, mas que tornam a resposta muito provável: Para um tópico como [Cantores vaiados nos festivais de música brasileira na década de 60], Chico Buarque é uma resposta correta, mas não há, na página Chico Buarque, nenhuma menção explícita à vaia, apenas o seguinte comentário:

   > *Mas desta vez a vitória foi contestada pelo público, que preferiu...*
   >
   > Chico Buarque
   > Wikipédia

   Portanto, de um ponto de vista lógico seria preciso mais informação para que uma dada resposta seja considerada correta, já que "ser contestada" não significa, necessariamente, "ser vaiada", mas a maior parte das pessoas consideraria a resposta certa e suficientemente justificada, visto que interpretariam "vaiado" não necessariamente de forma literal. Este é/seria um tipo de interpretação que sistemas automáticos teriam certamente dificuldade em fazer, mas que é rotineiramente realizado, inconscientemente, por seres humanos em comunicação.

De um outro ponto de vista, a diferença apontada também pode ser entendida como, de um lado, o que é considerado informação básica (pressuposta em uma pergunta, e portanto não necessariamente necessitando de explicitação... tal como Luanda ser capital de Angola; e, por outro, informação menos essencial, e que portanto deve ser justificada (no sentido de que esse é o objetivo não-trivial da pergunta). Ou seja, que entre as várias peças ou pepitas de informação, algumas são mais relevantes do que outras.

Isto pode ser, em termos puramente linguísticos, explicitado entre informação pressuposta pela pergunta e informação afirmada, ou melhor, requerida, pela pergunta.

Por exemplo, em "que capitais de língua portuguesa se canta o fado?" está pressuposto que a pessoa que pergunta sabe quais são as capitais de língua portuguesa e que pressupõe que a que responde também.

Por outro lado, existe ainda outra fonte de problemas, ou que requer clarificação. No Págico nós postulámos que as respostas deviam ser fundamentadas na wikipédia, melhor, na coleção que preparámos para o efeito (Simões, Costa e Mota, 2012).

No caso das justificativas decorrentes de um conhecimento estável, é sempre preciso imaginar que sistemas poderiam recorrer a bases de dados com conhecimento geográfico, por exemplo, para auxiliá-los na resposta, tal como as pessoas usariam a sua cultura geral. Por outro lado, lembramos que, no contexto do Págico, é importante que a resposta esteja fundamentada na Wikipédia, visto que outro dos objetivos do Págico era ver a que ponto a Wikipédia estava bem equipada.

Com isso, fica marcado de forma muito clara que, no âmbito do Págico, tão importante quanto oferecer uma página-resposta correta, é demonstrar que todo o "raciocínio" subjacente à resposta também está sustentado em informação da Wikipédia.

Essa é aliás uma das razões por que aceitamos a classificação de resposta certa mas não justificada, ou apoiada por páginas da wikipédia.

Seja como for, um trabalho que seria útil e interessante fazer era uma anotação das perguntas, e das respostas, em termos do pressuposto e do realmente perguntado, assim como das várias partes e/ou cadeias de inferência necessárias para chegar a uma resposta final correta e devidamente justificada. Veja-se (Santos et al., 2012) para uma proposta nesse sentido.

7 Observação dos comentários

Na tabela 1 apresentamos os tópicos a cujas respostas houve mais comentários (por parte dos avaliadores).

No entanto, o número de comentários por tópico não deve, por si só, ser considerado um indicador de dificuldade, visto muitos comentários terem a intenção de explicar o motivo da rejeição da resposta ou assinalar que não havia necessidade de justificação, e não refletem sempre dúvidas durante a avaliação.[4]

Em 31 tópicos (lista abaixo) (que originaram 61 comentários), o comentário apontava que a justificativa apresentada era desnecessária. Esta informação, embora não tendo sido levada em conta nas medidas de avaliação, é importante para esclarecer as possíveis razões da sua inclusão, por oposição aos casos em que a justificação é necessária.

Além disso, e embora tenhamos de reforçar que a questão dos comentários não foi sistemática, e portanto não deve ser demasiado levada a sério, podemos apresentar a lista de tópicos em que não houve comentários, assim como em que casos foi comentado que a resposta estaria nas caixas de informação (infoboxes) mas não na coleção do Págico.[5]

Seja como for, para uma próxima edição, ou se pudéssemos refazer todo o processo, deveríamos ter desenvolvido um sistema automático que permitisse ao avaliador, com um trabalho mínimo, escolher a causa da incorreção ou da dúvida, nos casos seguintes, que já sabemos serem possíveis e frequentes, e que poderíamos portanto ter tentado quantificar:

- resposta de um tipo diferente
- falta de justificativa parcial
- informação na wikipédia atual mas não na coleção
- necessidade de uma inferência adicional
- incerteza do avaliador

De qualquer modo, existiram casos ainda mais complexos, que passamos agora a discutir.

Como já mencionado, o **tópico 18** [Discos considerados marcantes...] trazia um dado subjetivo interessante em sua formulação: como "marcantes" seria interpretado pelos sistemas – e, mesmo, pelas pessoas? De fato, a análise dos comentários dos avaliadores mostra que esse foi o tópico que recebeu mais comentários. Marcante foi, principalmente, "traduzido" em termos de vendas, identificado em expressões como "mais vendido do Brasil", "disco de ouro"; "de diamante". Do lado dos avaliadores, os comentários revelam que o critério do número de vendas foi questionado, ou pelo menos não indiscutível:

[4] Além disso, quanto mais respostas, mais comentários possíveis, por isso o número de comentários por si só nunca podia ser uma medida, teria de ser pesado pelo número de respostas diferentes a esse tópico.

[5] Esta última questão apenas foi analisada/levada em conta (e, portanto, comentada) por um dos avaliadores, convém também dizer.

Tópico	Comentários
18 Discos brasileiros considerados marcantes na história da música brasileira	13
16 Membros da igreja associados à Teologia da Libertação	11
19 Tribos indígenas que vivem na Amazónia.	11
43 Produtos agrícolas com os quais se pode produzir combustível em escala comercial	11
150 Empresários lusófonos com uma fortuna considerável	11
9 Comidas de santo (...) que também fazem parte da culinária brasileira.	9
61 Movimentos culturais (...) que se refletiram nas artes plásticas e na música	9
13 Dinossauros carnívoros que habitaram o Brasil.	8

Tabela 1: Tópicos com mais comentários dos avaliadores

- *aceito que o disco mais vendido do Brasil seja marcante :(*;
- *Ser o album mais vendido não implica que tenha sido marcante na história da música brasileira*;
- *Vender muito é marcante?*

Certamente o questionamento se deve a alguns dos resultados obtidos segundo o critério das vendas, já que dificilmente alguém consideraria um disco como "Músicas para Louvar o Senhor" marcante na história da música brasileira, ainda que tenha vendido muito. Por outro lado e em outras épocas, o equivalente[6] pode ter sido de facto marcante na história da música, se pensarmos em obras de música sacra de Bach ou Handel. É apenas o nosso conhecimento factual da obra em questão que permite identificar que a causa da venda foi primordialmente religiosa e não (também) musical, e não algo que pudéssemos explicitar como uma regra sem exceções.

7.1 Exemplos de divergências entre os avaliadores

Conforme já explicado, os prazos apertados não nos permitiram uma avaliação sobreposta conforme o SIGA permite e era a nossa intenção efetuar. Contudo, os poucos casos em que houve sobreposição permitiram mesmo assim identificar alguns casos sobre os quais vale a pena refletir:

Lógica versus uso: No caso do [103]Movimentos culturais surgidos no nordeste do Brasil, a conceituação de "movimento cultural"levou a diferentes interpretações dos avaliadores. Assim, a resposta Manguebeat.683b02, com justificação "Cultura da região Nordeste do Brasil" e "Recife" foi considerada certa por um avaliador e errada por outro, com o argumento de que para ser cultural tem de ser mais do que musical (porque nesse caso se empregaria o termo musical e não cultural). Ninguém põe ou pôs em dúvida que a música é uma forma de cultura, mas sim se o termo "movimento cultural" se pode empregar para significar apenas "género musical". De um ponto de vista da classificação do Págico, aceitámos a resposta como correta – mas o que interessa é que este tipo de considerações são relevantes, e indiciam como as relações semânticas são fluidas e variáveis no uso.

Independentemente do veredito usado em relação à resposta concreta em causa, temos de salientar que, se um avaliador considerar que movimento musical não qualifica como movimento cultural, não irá ler com atenção o resto da página ou da resposta, e poderá portanto não reparar que isso está de facto indicado na página em questão:

> *Devido a principal bandeira do mangue ser a diversidade, a agitação na música contaminou outras formas de expressão culturais como o cinema, a moda e as artes plásticas.*
>
> mangue
> Wikipédia

Da mesma forma, outra discussão que surgiu é o verdadeiro significado de "filmes sobre um determinado assunto ou tema", que demonstra muito claramente como há ou pode haver graus de correção numa resposta.

Assim, no [Filmes sobre o cangaço], chegámos à conclusão de que existe uma clara ordem decrescente entre documentários, filmes históricos, filmes com um enredo em que o cangaço é proeminente, até porno-chachadas ou filmes pornográficos tendo como ambiente elementos dessa realidade. Onde dividir? Aceitar tudo, ou apenas filmes que pudessem descrever-se naturalmente em português como "filmes sobre o cangaço"? Por um lado, tal depende da intenção do perguntador... se estivesse interessado em estudar a influência dessa questão na cinemator-

[6] No sentido de música composta com intenção religiosa, ou seja, de louvar o senhor, aumentar o sentimento religioso dos ouvintes, ser apropriada para ouvir em cerimónias religiosas.

grafia brasileira, provavelmente todos os filmes teriam (até igual) interesse. Se por outro lado fosse um historiador ou um aluno que estava interessado em história, apenas os primeiros da lista seriam de apresentar. Este é um caso onde nos parece claro que existe ordem de topicalidade da resposta que seria extremamente útil conseguir codificar e apresentar. Ou seja, mais importante que decidir qual a linha de demarcação, apresentar casos indiscutíveis e outros mais periféricos, como tal.

O mesmo caso, de uma gradação que em última análise se terá sempre de considerar subjetiva, aconteceu nos casos de [filmes sobre futebol], em que os diferentes avaliadores usaram estratégias ou critérios diferentes para decidir, não aceitando que bastaria que na sinopse do filme houvesse menção a futebol para a resposta dever ser considerada correta. Vejamos exemplos concretos:

Na sinopse de um dos filmes, a única menção a futebol era:

> *Entre os seus alunos estão Acácio, um jogador de futebol que está prestes a se mudar para a Inglaterra, (...)*
>
> Wikipédia

mas tal pareceu suficiente para que o avaliador considerasse a resposta correta, comentando "não é sobre futebol, mas o futebol parece ser parte importante...". Na página de outro filme, a única menção a futebol informava que

> *Ainda criança Dé vê seu irmão ser assassinado por um traficante por conta de uma briga num jogo de futebol.*
>
> Era uma Vez... filme
> Wikipédia

e isso foi considerado dado insuficiente para que a resposta fosse aceita, como o comentário ilustra: "Embora haja alguma coisa com futebol no filme, recuso-me a considerar o Romeu e Julieta um filme sobre futebol". Em conclusão, os avaliadores tiveram pontos de vista divergentes, e mesmo que todos os avaliadores tivessem avaliado todas as respostas e todas as discordâncias tivessem sido resolvidas por maioria, isso não garantia que a avaliação, mesmo que fosse mais consistente, fosse representante da verdade ou mesmo da opinião dos participantes.

8 Comentários finais

Esperamos ter demonstrado que a avaliação de respostas ao Págico não é uma mera questão de sim ou não.

Pelo contrário, existem diversos eixos que permitem uma diversificação do grau de resposta: o grau de conhecimento partilhado (e assumido) entre a pessoa que perguntou e quem responde; algo ser útil embora não diretamente uma resposta completa; ou respostas que apenas fazem sentido em determinados contextos.

Além disso, não há critérios, na maior parte dos casos, que sejam tão específicos que não aceitem interpretações mais alargadas, ou inferências que não tenham escapado ao organizador mais prevenido – uma simples leitura dos artigos dos participantes, neste volume, e em particular das questões ou tópicos que eles apresentam como problemáticos ou mal definidos, dá-nos imediatamente razão.

Neste artigo, por isso, além de documentar o que fizemos no Págico, tentámos generalizar a experiência apontando alguns problemas que na nossa opinião se põem em qualquer trabalho que tem a ver com o uso da língua.

Em última análise, insistimos que não é possível, nem interessante, ser mais rigoroso do que a própria língua, e que portanto devemos aceitar que existem várias interpretações possíveis, e várias formas de enriquecer um dado assunto ou pergunta. Perguntas autênticas (e não as de jogos em que só há uma resposta certa, e que foram fixadas na área de resposta automática a perguntas (RAP) com o nome de "factóides" (Magnini et al., 2005)) implicam algum desconhecimento da parte de quem pergunta, com a consequente humildade de aceitar várias respostas e várias informações colaterais como parte integrante do processo de aprendizagem.

Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN, e, durante 2011, pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN).

Agradecemos ao resto da organização do Págico, sem a qual o mesmo não teria sido possível, e a todos os participantes, cujas respostas ajudaram a iluminar os tópicos e a esclarecer pontos pouco claros.

Estamos também gratos à Stella Tagnin e ao Alberto Simões pela revisão feita, que

nos permitiu tornar o artigo mais legível e esclarecedor.

Referências

Freitas, Cláudia. 2012. A lusofonia na wikipédia em 150 tópicos. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Magnini, Bernardo, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov, e Richard Sutcliffe. 2005. Overview of the CLEF 2004 Multilingual Question answering track. Em Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, e Bernardo Magnini, editores, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science*, pp. 371–391, Berlim/Heidelberg. Springer.

Mota, Cristina. 2012. Resultados págicos: participação, medidas e pontuação. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Rocha, Paulo e Diana Santos. 2007. CLEF: Abrindo a porta à participação internacional em avaliação de RI do português. Em Diana Santos, editor, *Avaliação Conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, capítulo 13, pp. 143–158.

Santos, Diana. 2007. Evaluation in natural language processing, 6-17 Agosto, 2007. Curso na ESSLLI 2007, European Summer School on Language, Logic and Information ESSLLI, Dublin, Irlanda, `http://www.linguateca.pt/Diana/download/EvaluationESSLLI07.pdf`.

Santos, Diana. 2012. Translation. Em Robert Binnick, editor, *Handbook of Tense and Aspect*. Oxford University Press.

Santos, Diana e Luís Miguel Cabral. 2009. Summing GikiCLEF up: expectations and lessons learned. Em Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, Giovanna Roda, Francesca Borri, Alessandro Nardi, e Carol Peters, editores, *Multilingual Information Access Evaluation, Vol. I: Text Retrieval Experiments*, volume Vol. I: Text Retrieval Experiments, pp. 212–222, Berlim / Heidelberg. Springer.

Santos, Diana, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Julia Schulz, Yvonne Skalban, e Erik Tjong Kim Sang. 2010. GikiCLEF: Crosscultural Issues in Multilingual Information Access. Em Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, Maio, 2010. European Language Resources Association (ELRA).

Santos, Diana, Cristina Mota, Alberto Simões, Luís Costa, e Cláudia Freitas. 2012. Balanço do Págico e perspetivas de futuro. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Simões, Alberto, Luís Costa, e Cristina Mota. 2012. Tirando o chapéu à Wikipédia: A coleção do Págico e o Cartola. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Simões, Alberto, Paulo Rocha, e Rúben Fonseca. 2009. Webpaper — más perguntas e boas respostas: a arte de interrogar. Em Paulo Dias, António José Osório, e Altina Ramos, editores, *O digital e o currículo*. Centro de Competência da Universidade do Minho, pp. 227–238, Maio, 2009.

Sparck Jones, Karen. 2003. Is question answering a rational task? Em R. Barnardi e M. Moortgat, editores, *Questions and Answers: Theoretical and Applied Perspectives, Second CoLogNET-ElsNET Symposium*. Utrecht Institute of Linguistics, pp. 24–35.

Voorhees, Ellen M. e Dawn M. Tice. 2000. Building a Question Answering Test Collection. Em Nicholas Belkin et al, editor, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 200–207, Atenas, 24-28 Julho, 2000.

Resultados págicos: participação, medidas e pontuação

Cristina Mota
Linguateca/FCCN
cmota@ist.utl.pt

Resumo

Este artigo descreve a participação no Págico, tanto a nível de sistemas, como a nível de participação humana. Além disso, caracteriza o processo de avaliação e apresenta as medidas de avaliação implementadas, introduzindo as novas medidas de pseudo-abrangência, pseudo-medida-F, originalidade e criatividade. Finalmente, mostra os resultados globais por participante em vários cenários de avaliação, bem como os resultados detalhados por temas e lugares dos tópicos no cenário completo do Págico, contrastando a participação humana e a automática.

Palavras chave

Recolha de informação, Resposta a perguntas, Avaliação, Cooperação pessoa-máquina, Wikipédia

1 Apresentação

O Págico foi uma avaliação conjunta em que sistemas e pessoas tiveram de fornecer respostas a 150 tópicos (consulte-se (Mota et al., 2012) para uma descrição da tarefa, e (Santos, 2012) para uma motivação da avaliação, e (Freitas, 2012) para uma apresentação e discussão do processo de criação dos tópicos). As respostas deveriam ser encontradas numa versão estática da Wikipédia portuguesa (veja-se (Simões, Costa e Mota, 2012) para uma descrição e avaliação deste recurso) e correspondiam aos títulos das páginas da Wikipédia. Nos casos em que o conteúdo da página não justificava só por si que a página (ou seja, o seu título) era a resposta correta, os participantes tinham de fornecer adicionalmente a(s) página(s) que permitia(m) chegar a essa conclusão - designaremos o conjunto das justificações simplesmente por justificação.

A título de exemplo, para um tópico como [Que cientistas ou avanços da ciência podem ser direta ou indiretamente relacionados com os jesuítas da escola de Coimbra?] esperava-se que os participantes identificassem *Nónio* (Wikipédia), entre outras respostas possíveis. Nesse caso, como a página não contém informação suficiente que justifique que o nónio foi um avanço relacionado com os jesuítas de Coimbra, os participantes tinham de associar a essa resposta a página *Pedro_Nunes_(matemático)* (Wikipédia), que era igualmente uma boa resposta, como justificação, pois é ela que contém essa informação.[1]

Sistemas e participantes humanos participaram no Págico de forma distintas. Enquanto os primeiros foram buscar os recursos da avaliação fornecidos pela organização (versão estática da Wikipédia portuguesa e lista de tópicos de avaliação) para os processarem e depois enviarem as respostas num ou mais ficheiros (até um máximo de três) que designaremos *corridas*, os participantes humanos forneceram as respostas através de uma interface desenvolvida para o efeito no SIGA (Costa, Mota e Santos, 2012) que permitia fazer pesquisas na versão estática da Wikipédia e adicionar as páginas como resposta ou justificação.

O calendário para participação também não foi o mesmo, uma vez que se deu mais tempo aos participantes humanos para responderem aos tópicos. O período para envio de respostas de participantes humanos e sistemas teve início ao mesmo tempo, a 4 de Novembro de 2011, e decorreu até 11 de Novembro para sistemas e 30 de Novembro para participantes humanos.

Após ter fechado o período de envio de corridas por sistemas[2], iniciou-se o processo de avaliação humana das respostas. Os primeiros resultados foram divulgados a 9 de Janeiro de 2012, quando mais de metade das respostas já tinham sido avaliadas por avaliadores humanos e a 21 de Janeiro concluiu-se a avaliação das respostas por pelo menos um avaliador humano. Os resultados finais, em que algumas das respostas foram revistas por mais de um avaliador e em

[1] Este exemplo foi retirado da página de apresentação e divulgação do Págico em `http://www.linguateca.pt/Pagico/`.

[2] Além das corridas oficiais, demos a possibilidade de serem enviadas corridas não oficiais pelos sistemas, após o prazo final. Essas corridas foram avaliadas automaticamente, mas neste artigo esses resultados não serão tidos em conta.

que foram revistas as justificações dadas pelos criadores dos tópicos, foram divulgados a 18 de Fevereiro.

Este artigo foca a participação no Págico, contrastando sistemas e participantes humanos, o processo geral de avaliação e as medidas utilizadas para avaliar as respostas dos participantes. Além de mostrar, por participante, os resultados globais da avaliação, mostra ainda esses resultados detalhados por tema, subtema e localização dos tópicos. Como, mais do que saber quem é o melhor participante, se pretende perceber em que difere a participação humana da automática, compara ainda os resultados entre ambas.

2 Participação no Págico

Inscreveram-se no Págico 21 equipas: 6 sistemas e 15 participantes humanos. No entanto, apenas um terço dos inscritos acabou por participar efetivamente: 2 sistemas e 5 participantes humanos. É notável também que pouco mais de um terço (4 sistemas e 4 participantes humanos) desistiram de participar sem sequer ver a coleção com os tópicos de avaliação `Págico`, embora dois sistemas e um dos participantes humanos ainda tenham visto a coleção com exemplos de tópicos, `PagicoEXEMPLOS`. Finalmente, pouco menos de um terço (os restantes 6 participantes humanos) desistiram depois de terem visto a coleção `Págico` e de terem feito algumas pesquisas na coleção, sendo que três deles ainda visualizaram documentos e chegando um deles a responder a dois tópicos. A tabela 1 sintetiza o perfil de envolvimento das equipas que se inscreveram no Págico.

Apresentamos sucintamente, em seguida, os sete participantes que forneceram respostas:

GLNISTT 23 estudantes organizados em 8 grupos, em que cada estudante respondeu em média a 7 perguntas. Os grupos responderam a um conjunto de tópicos disjuntos. Participaram no âmbito de um projecto para a cadeira de Língua Natural, de mestrado, sendo o principal objectivo perceber o que seria necessário fazer para construir um sistema capaz de participar no Págico. A participação dos 8 grupos foi avaliada como um todo, mas também individualmente por grupo, tendo sido facultado esse resultado à professora responsável pela cadeira. Recorreram a várias fontes, incluindo a Wikipédia atual. (Coheur e Ângela Costa, 2012)

ludIT Equipa de 6 pessoas que também se organizaram de modo a responderem a conjuntos de tópicos disjuntos. No entanto, colaboraram entre si em caso de dúvida. Usaram uma estratégia de pesquisa na Wikipédia atual e confirmação das respostas na versão da Wikipédia usada no Págico. (Veiga et al., 2012)

João Miranda Participou individualmente e usou uma estratégia de pesquisa com base nos termos do tópico ou tentando procurar pelo nome do artigo que contém a resposta no caso do tópico lhe ser familiar. (Miranda, 2012)

Ângela Mota; Bruno Nascimento Participaram individualmente, mas não enviaram relatos da participação.

RAPPORTAGICO Este sistema combina o reconhecimento de sintagmas frásicos com a identificação de sinónimos recorrendo a uma ontologia lexical. Enviou três corridas, em que uma delas serve de base de comparação (em inglês, *baseline*) às outras duas, que fazem expansão de sinónimos dos sintagmas verbais, cada uma com métodos de expansão diferentes: uma com *Bag of Words* e a outra com *Personalized Page Rank*. (Rodrigues, Gonçalo Oliveira e Gomes, 2012)

RENOIR Usou um sistema de recuperação geográfica que devolve os documentos mais relevantes para os tópicos, os quais numa das corridas não foram reformulados nem lematizados, numa outra foram lematizados mas sem que tivessem sido reformulados e na terceira foram lematizados e reformulados. (Cardoso, 2012)

A tabela 2 mostra informações sobre as respostas fornecidas pelos participantes. A distinção mais evidente é que os sistemas não forneceram justificações adicionais para as respostas, além de terem enviado, como se esperaria, um maior número de respostas. Repare-se também que os participantes humanos repetiram entre si menos respostas do que os participantes automáticos, embora se deva salientar desde já que apenas dois participantes humanos responderam a 148 ou mais tópicos, tendo um deles respondido a todos, e que os outros três responderam no máximo a um terço dos tópicos cada um. Sobre esta questão de a quantos tópicos os participantes responderam, bem como do número médio de respostas por tópico dadas pelos participantes, consulte-se a tabela 6, na secção 5.

		Tipo de participação	
		Automática	Humana
Inscritos		6	15
- Participantes		**2**	**5**
- Desistentes	responderam a tópicos	-	1
	consultaram páginas	-	3
	fizeram pesquisas sem consultar	-	2
	viram apenas exemplos	2	1
	não viram coleções	2	3

Tabela 1: Perfil de envolvimento no Págico.

Tipo de Participação	Participante (Corrida)	# Respostas	# Com justificação
Humana	Ângela Mota	157	8 (5%)
	GLNISTT	1016	255 (25%)
	ludIT	1387	489 (35%)
	João Miranda	101	60 (50%)
	Bruno Nascimento	34	1 (3%)
	Total	**2695**	
	Distintas	**2383**	
	Total/Distintas	**1.13**	
Automática	RENOIR (1)	15000	
	RENOIR (2)	15000	
	RENOIR (3)	15000	
	Total	**45000**	
	Distintas	**28626**	
	Total/Distintas	**1.57**	
	RAPPORTAGICO (1)	1718	
	RAPPORTAGICO (2)	1736	
	RAPPORTAGICO (3)	1730	
	Total	**5184**	
	Distintas	**2343**	
	Total/Distintas	**2.21**	
	Total	**50184**	
	Distintas	**30543**	
	Total/Distintas	**1.64**	
Total		**52879**	
Distintas		**32485**	
Total/Distintas		**1.62**	

Tabela 2: Participantes no Págico.

3 Procedimento de avaliação

Como referido anteriormente, na secção 1, sistemas e participantes humanos forneceram as respostas de modo distinto: os primeiros enviaram ficheiros com as respostas e os segundos utilizaram a interface do SIGA dedicada a esse fim.

No entanto, o procedimento de avaliação, adaptado do GikiCLEF (Santos et al., 2010), que não teve participação humana, não fez distinção entre os dois tipos de participação, a não ser durante a avaliação humana em que foram apresentados aos avaliadores, caso existissem, os comentários dados pelos participantes para complementar a justificação.

A avaliação processou-se então em cinco passos, que se descreverão em seguida: geração do monte (em inglês, *pool*) das respostas, avaliação automática das respostas, distribuição pelos avaliadores das respostas a avaliar, avaliação humana das respostas e cálculo das medidas de avaliação.

3.1 Geração do monte das respostas

O monte corresponde à união de todas as respostas, ou seja, ao conjunto de respostas distintas dadas por todos os participantes. É também criada uma lista com todas as respostas fornecidas pelos participantes. Antes da geração do monte, para cada participante humano é gerada uma corrida com as suas respostas, cujo formato é idêntico ao das corridas dos sistemas. Assim, as respostas dos sistemas e dos participantes humanos serão tratadas de forma

indistinguível. Em (Costa, Mota e Santos, 2012) justifica-se e descreve-se em maior detalhe esta opção.

Como já se viu na tabela 2, a partir das 52879 respostas enviadas, foi criado um monte com 32485 respostas, as quais foram então avaliadas. De notar que se não tivermos em conta as justificações, então o número de respostas distintas é 32086, o que quer dizer que 399 casos correspondem a respostas idênticas mas com justificações diferentes.

3.2 Avaliação automática das respostas

Após o monte ter sido gerado, as respostas nele contidas (32485) foram avaliadas automaticamente, comparando as respostas dos participantes e as fornecidas pelos criadores dos tópicos da seguinte forma:

- todas as respostas que correspondam a documentos inválidos (do ponto de vista de não poderem ser respostas no Págico), como sejam, páginas de categorias, de predefinição, de desambiguação, figuras, MediaWiki, ficheiros ou portal são avaliadas automaticamente como incorretas e não passarão para a fase de avaliação humana (4292 respostas);
- se o par (resposta, justificação) tiver sido fornecido pelos criadores dos tópicos, a resposta é considerada correta e justificada e será avaliada automaticamente como correta (420 respostas);
- se a resposta tiver sido fornecida pelos criadores dos tópicos, mas a justificação não, então a resposta é considerada correta, mas não justificada, e, como tal, a resposta será avaliada automaticamente como incorreta; este par passa para a fase de avaliação humana, de forma a validar se de facto a resposta é ou não justificada (235 respostas);
- se a resposta não tiver sido fornecida pelos criadores dos tópicos, então a resposta é avaliada automaticamente como incorreta; o par passa para a fase de avaliação humana, de forma a validar se a resposta é ou não correta e se está ou não justificada (27536 respostas).

A tabela 3 mostra o resultado da avaliação automática, contrastando a participação automática e humana. Salienta-se que a maioria das respostas avaliadas automaticamente como corretas e justificadas foram dadas exclusivamente por participantes humanos (58%), mas que uma parte significativa foi mesmo assim dada tanto por participantes humanos como sistemas (31%) e que, além disso, 11% foram dadas exclusivamente por sistemas. Também se pode ver que praticamente todas as respostas avaliadas automaticamente como corretas, mas cuja justificação não foi idêntica à dos criadores dos tópicos, foi dada por participantes humanos. No caso dos participantes humanos isso pode querer dizer que (i) a página dada pelo participante como justificação é diferente da usada pelo criador do tópico, (ii) o participante forneceu mais páginas de justificação além das dadas pelo criador do tópico, que pode até ter considerado a resposta como auto-justificada, e (iii) o participante não forneceu página de justificação quando o criador do tópico estabeleceu justificação adicional; nos casos as que as respostas dos sistemas foram consideradas corretas, mas não justificadas, são casos como em (iii), pois os sistemas não forneceram quaisquer justificações adicionais (cf. tabela 2).

As tabelas 11, 12 e 13, na secção 5, ilustram, respetivamente, os tópicos com mais respostas corretas e justificadas dadas exclusivamente por participantes humanos, exclusivamente por automáticos e por ambos, após concluída toda a avaliação (automática e humana).

Vale a pena referir que embora o número de respostas predeterminadas pelos criadores dos tópicos tenha acabado por ser inferior ao número de respostas corretas encontradas pelos participantes humanos, das 708 respostas definidas previamente pelos criadores dos tópicos, 288 não foram dadas por qualquer participante. No entanto, se não se tiver em conta as justificações a elas associadas, então o número de respostas não encontradas pelos participantes desce para 184. Isto quer dizer que 104 respostas dadas pelos criadores dos tópicos foram também usadas pelos participantes, mas estes não as justificaram da mesma forma.

3.3 Distribuição pelos avaliadores das respostas a avaliar

Avaliador	# Todas	# Só humanas
Cláudia	653	182
Cristina	279	265
Diana	570	121
Luís	818	299
Paulo	25896	1464

Tabela 4: Distribuição das avaliações humanas.

As respostas foram inicialmente distribuídas de forma disjunta pelos avaliadores e 266 respos-

Avaliação	Humanas	Automáticas	Ambas	Total
Correta e justificada	243	48	129	420
Correta e não justificada	221	8	6	235
Incorreta (porque documento inválido)	0	4292	0	4292
Restantes casos	1477	25753	306	27536
Total	1941	30101	441	32483

Tabela 3: Estatísticas da avaliação automática.

tas, todas avaliadas pelo mesmo avaliador, foram depois atribuídas a um segundo avaliador de um grupo de três avaliadores. A tabela 4 mostra que quase 92% das respostas foram avaliadas pelo mesmo avaliador, mas que se excluirmos as respostas fornecidas exclusivamente por sistemas, então a fatia avaliada por esse avaliador reduz para 62%.

3.4 Avaliação humana das respostas

Após a avaliação automática, as respostas que foram avaliadas automaticamente como corretas mas não justificadas, e como incorretas (por essa resposta não ter sido dada pelos criadores de tópicos[3]), foram alvo de avaliação por avaliadores humanos. Esta avaliação engloba os seguintes passos:

- Avaliação das respostas por avaliadores humanos. A avaliação humana foi feita através do SIGA. Cada avaliador teve acesso à avaliação automática das respostas que lhe foram atribuídas. Além de poder considerar a resposta como correta ou incorreta, podia deixar a resposta por avaliar ou considerá-la duvidosa; caso a considerasse correta (ou a resposta já tivesse sido avaliada automaticamente com correta), tinha também de julgar se estava ou não justificada. Além de avaliarem as respostas e justificações, os avaliadores podiam associar comentários a cada avaliação que fizeram.
- Resolução de conflitos e revisão. Casos duvidosos foram sendo discutidos durante e após a avaliação.

Em (Freitas et al., 2012) descreve-se em maior detalhe a avaliação humana, apresentando-se os critérios para considerar ou não uma resposta como correta, e discutindo-se as várias dificuldades envolvidas nesta fase.

[3]Como referido antes, respostas que não foram dadas pelos criadores de tópicos, mas que correspondem a documentos inválidos - páginas de desambiguação, redireção, predefinição ou com conteúdo audiovisual, - são automaticamente consideradas incorretas, mas não passam para a avaliação humana.

3.5 Cálculo das medidas de avaliação

Este passo consiste no cálculo de cada uma das medidas de avaliação de acordo com a avaliação feita para cada resposta. Foram calculadas as seguintes medidas, descritas na secção 4: precisão, precisão tolerante, pseudo-abrangência, pseudo-medida-F, originalidade, criatividade e pontuação final.

Dado que aos participantes humanos foi dada a possibilidade de responderam a um subconjunto dos tópicos de avaliação por si escolhidos de entre os 150, seguimos a tradição das avaliações da Linguateca de fazer a avaliação dos participantes por cenários (cf. (Costa, Rocha e Santos, 2007; Gonçalo Oliveira et al., 2008; Oliveira et al., 2008)).

Em particular, adoptámos a avaliação por cenários tal como definida no Segundo HAREM, em que cada participante é avaliado no seu cenário, bem como em todos os outros, incluindo o do Págico que é constituído por todos os tópicos. Isso permite comparar participantes que responderam a um subconjunto dos tópicos de avaliação com os restantes participantes que responderam a todos ou a um outro subconjunto de tópicos. Para tal, no cálculo das medidas de avaliação ignora-se das corridas desses participantes as respostas a tópicos que não pertencem a esse subconjunto.

Cenário	# Tópicos
Págico/ludIT	150
GLNISTT	148
Ângela Mota (AM)	50
João Miranda (JM)	40
Bruno Nascimento (BN)	18

Tabela 5: Cenários do Págico.

No Págico, um cenário é então definido por um conjunto de tópicos. Além do cenário Págico, constituído pelos 150 tópicos, criámos um cenário por cada participante que respondeu a um subconjunto desse e que é constituído por esse subconjunto de tópicos. Uma vez que o participante ludIT respondeu a todos os tópicos, o seu cenário é igual ao cenário Págico. A tabela 5 mostra por quantos tópicos é constituído cada cenário.

4 Medidas de avaliação

Os participantes foram avaliados no Págico de acordo com as medidas de avaliação usadas no GikiCLEF (precisão e pontuação final), e também com as seguintes novas medidas: pseudo-abrangência, pseudo-medida-F, originalidade e criatividade.

Essas medidas foram calculadas para cada corrida, e as medidas de originalidade e criatividade foram também calculadas para cada participante. Neste último caso, as diferentes corridas de um mesmo participante foram vistas como uma única corrida.

Cada uma das medidas será descrita em seguida, usando a seguinte notação e terminologia:

p = participante p

c = corrida c

C = conjunto das respostas correctas e justificadas corretamente

$\tilde{C}$ = conjunto das respostas correctas e justificadas incorrectamente

R = conjunto das respostas fornecidas pelos participantes

T = conjunto de tópicos

Designaremos simplesmente por *resposta* o par composto pela resposta e a sua justificação, uma vez que geralmente, em contexto, não é ambígua a sua interpretação. Assim, entende-se por *resposta correta* uma resposta que está correta e a sua justificação está correta também. Quando a resposta não foi justificada corretamente é designada por *resposta não justificada*.

4.1 Precisão

$$P_{p,c} = \frac{|C_{p,c}|}{|R_{p,c}|} \quad (1)$$

A precisão $P_{p,c}$ é uma medida que avalia a qualidade das respostas e respetivas justificações incluídas na corrida c do participante p, e é dada pela fórmula 1, em que $R_{p,c}$ e $C_{p,c}$ são, respetivamente, o número de respostas dadas e de respostas corretas c do participante p.

4.2 Precisão tolerante

$$\tilde{P}_{p,c} = \frac{|C_{p,c}| + |\tilde{C}_{p,c}|}{|R_{p,c}|} \quad (2)$$

A precisão tolerante $\tilde{P}_{p,c}$ é uma variante da medida de precisão $P_{p,c}$ que avalia a qualidade das respostas incluídas na corrida c do participante p sem ter em conta a correção das justificações. $\tilde{P}_{p,c}$ é então dada pela fórmula 2, em que $R_{p,c}$, $C_{p,c}$ e $\tilde{C}_{p,c}$ são, respetivamente, o número de respostas dadas, de respostas corretas e de respostas corretas e não justificadas da corrida c do participante p.

4.3 Pseudo-abrangência

$$\alpha_{p,c} = \frac{|C_{p,c}|}{|C_{Pagico}| + |C_{aval}|} \quad (3)$$

Quando se conhece à partida todas as respostas corretas que um participante deve fornecer, é usual calcular uma medida de abrangência que avalia a quantidade de respostas que o participante forneceu relativamente ao que devia ter fornecido. Esse é o caso, por exemplo, em avaliações de reconhecimento de entidades mencionadas em que os textos anotados pelos participantes são comparados aos mesmos textos em que as entidades mencionadas a reconhecer foram exaustivamente anotadas pela organização da avaliação. O conjunto desses textos é designado por coleção dourada. Veja-se, por exemplo, (Gonçalo Oliveira et al., 2008) para uma descrição da abrangência usada no HAREM.

Em avaliações em que as respostas relevantes não se conhecem à partida, como acontece, por exemplo, em recolha de informação, em que não são conhecidos os documentos relevantes que os sistemas devem encontrar, a abrangência é calculada com base nos documentos relevantes conhecidos (**?**). Esses documentos são identificados por avaliadores humanos no monte dos documentos que é criado após todos os participantes terem enviado as suas corridas.

No Págico, calculámos uma variante da medida de abrangência a que chamámos *pseudo-abrangência* que tem em conta não só as respostas definidas à partida, mas também as respostas identificadas por avaliadores humanos como corretas no monte das respostas. Assim, a *pseudo-abrangência* $\alpha_{p,c}$ é dada pela fórmula 3 e calcula a quantidade de respostas corretas fornecidas pela corrida c do participante p relativamente ao total de respostas conhecidas no Págico, ou seja, relativamente ao total de respostas corretas fornecidas pelos criadores do tópicos, C_{Pagico}[4], juntamente com as respostas fornecidas por todos os participantes e que foram avaliadas como corretas, C_{aval}, e que não existem em C_{Pagico}.

4.4 Pseudo-medida-F

$$\phi_{p,c} = 2 \times \frac{P_{p,c} \times \alpha_{p,c}}{P_{p,c} + \alpha_{p,c}} \quad (4)$$

Em avaliações em que se calcula a precisão e a abrangência, também se costuma calcular

[4]Em alguns casos, os criadores dos tópicos forneceram respostas não justificadas que não são tidas em conta no cálculo da pseudo-abrangência.

a medida-F que combina as duas medidas anteriores num só valor.

No Págico, dado que temos precisão e pseudo-abrangência, calculámos a pseudo-medida-F, dada pela fórmula 4.

4.5 Originalidade

$$O_{p,c} = \sum_{i}^{T} \sum_{j}^{R_{p,c,i}} o(r_{p,c,i,j}) \tag{5}$$

$$o(r_{p,c,i,j}) = \begin{cases} p(i) & r_{p,c,i,j} \in C_{aval} \quad \wedge \\ & r_{p,c,i,j} \notin C_{Pagico} \quad \wedge \\ & r_{p,c,i,j} \notin \bigcup_{m \neq p, n \neq c} R_{m,n} \\ 0 & \text{c.c.} \end{cases} \tag{6}$$

No Págico definimos uma medida de originalidade por corrida, $O_{p,c}$, dada pela fórmula 5, que contabiliza o número de respostas corretas e originais da corrida c do participante p, ou seja, o número de respostas corretas que existem exclusivamente nessa corrida e que também não pertencem ao conjunto de respostas fornecidas pelos criadores dos tópicos. Uma resposta é tão mais original quanto maior for o número de participantes $p(i)$ que tentaram responder ao tópico i, como se pode ver pela fórmula 6, que calcula a originalidade da resposta j ao tópico i da corrida c do participante p, $o(r_{p,c,i,j})$.

$$O_{p} = \sum_{i}^{T} \sum_{j}^{R_{p,i}} o(r_{p,i,j}) \tag{7}$$

$$o(r_{p,i,j}) = \begin{cases} p(i) & r_{p,i,j} \in C_{aval} \quad \wedge \\ & r_{p,i,j} \notin C_{Pagico} \quad \wedge \\ & r_{p,i,j} \notin \bigcup_{m \neq p} R_{m} \\ 0 & \text{c.c.} \end{cases} \tag{8}$$

Nos casos em que o participante tem mais de uma corrida, a mesma resposta correta em corridas diferentes não é contabilizada como resposta original, mesmo que só tenha sido dada por esse participante. Por essa razão definimos ainda a originalidade por participante, O_p, dada pela fórmula 7, em que todas as corridas desse participante constituem uma só corrida.

Repare-se que tanto para se calcular $O_{p,c}$ como O_p a originalidade de uma resposta é proporcional a $p(i)$, ou seja, ao número de participantes que tentaram responder ao tópico. No caso de $O_{p,c}$, se a originalidade fosse proporcional ao número de corridas que tentaram responder ao tópico, estar-se-ia a penalizar os participantes (automáticos) que enviaram mais do que uma corrida, e entre as quais é natural que haja respostas repetidas.

4.6 Criatividade

$$K_{p,c} = \sum_{i}^{T} \sum_{j}^{R_{p,c,i}} k(r_{p,c,i,j}) \tag{9}$$

$$k(r_{p,c,i,j}) = \begin{cases} \frac{1}{c(r_{p,c,i,j})} \times p(i) & r_{p,c,i,j} \\ & \in C_{Pagico} \bigcup C_{aval} \\ 0 & \text{c.c.} \end{cases} \tag{10}$$

$p(i)$ = número de participantes no tópico i
$c(r_{p,c,i,j})$ = número de participantes que deram a resposta $r_{p,c,i,j}$

Uma resposta correta de um participante pode não ser original, por existir no conjunto de respostas determinadas pelos criadores dos tópicos ou por ter sido dada por mais do que um participante. No entanto, pode ser mais ou menos criativa, no sentido de haver menos ou mais participantes a darem a mesma resposta.

Definimos então uma medida de criatividade por corrida $K_{p,c}$, dada pela fórmula 9, que contabiliza quão criativas são as respostas da corrida c do participante p. A criatividade $k(r_{p,c,i,j})$ de uma resposta i ao tópico j da corrida c do participante p é inversamente proporcional ao número de participantes que deram a mesma resposta, $c(r_{p,c,i,j})$, e diretamente proporcional ao número de participantes que tentaram responder ao tópico, $p(i)$, tal como se pode ver na fórmula 10.

$$K_{p} = \sum_{i}^{T} \sum_{j}^{R_{p,i}} k(r_{p,i,j}) \tag{11}$$

$$k(r_{p,i,j}) = \begin{cases} \frac{1}{c(r_{p,i,j})} \times p(i) & r_{p,i,j} \\ & \in C_{Pagico} \bigcup C_{aval} \\ 0 & \text{c.c.} \end{cases} \tag{12}$$

À semelhança do que acontece na originalidade por corrida, em que respostas dadas unicamente por um único participante não contribuem para a originalidade da corrida se ocorrerem em mais do que uma corrida do mesmo participante, na criatividade por corrida, a criatividade de uma resposta é menor se tiver sido dada por corridas diferentes de um mesmo participante. Isso penaliza não só a criatividade

das corridas desse participante, mas também a das corridas de outros participantes que deram a mesma resposta.

Assim, definimos também a criatividade por participante, K_p, dada pela fórmula 11, que considera para cada participante a união de todas as suas corridas, e em que a criatividade $k(r_{p,c,j})$ é também proporcional ao número de participantes $p(i)$ que tentaram responder ao tópico i.

4.7 Pontuação final no Págico

$$M_{p,j} = |C_{p,c}| \times P_{c,j} \quad (13)$$

Embora tenhamos definido várias medidas para avaliar as corridas de perspectivas diferentes, os participantes foram classificados no Págico de acordo com a medida de classificação final por língua do GikiCLEF, que aqui designaremos $M_{p,c}$ e que é dada pela fórmula 13.

Esta medida, baseada na precisão $P_{p,c}$, permite distinguir participantes que tenham a mesma precisão com um número de respostas corretas diferentes. Mais especificamente, nessa situação, a medida atribui uma melhor pontuação final aos participantes que tenham mais respostas corretas.

5 Pontuação no Págico

Verificar se participantes humanos encontrariam mais respostas corretas na Wikipédia do que sistemas, ou se teriam um melhor desempenho do que sistemas nessa tarefa, não era um dos objectivos do Págico. Partiu-se do princípio de que encontrariam e de que seriam melhores. No entanto, comprova-se isso mesmo pelas tabelas 6 e 7: a primeira contém diversas estatísticas sobre as respostas ($|T|$ é número de tópicos respondidos, $|R|$ é o número de respostas dadas, $|R|/|T|$ é o número médio de respostas dadas por tópico, $|C|$ é o número de respostas corretas e bem justificadas e $\tilde{C}$ é o número de respostas corretas mas não justificadas corretamente; a segunda apresenta a avaliação dos participantes no Págico para cada uma das várias medidas de avaliação apresentadas nas secção 4. Ambas as tabelas mostram os valores calculados em cada um dos cenários do Págico.

Como seria de esperar, os participantes humanos tiveram uma precisão melhor do que os sistemas, acima dos 56% indo até quase 90% enquanto a dos sistemas não passou de 12%. O que talvez já não seja tão expectável é que os participantes humanos também acabaram por fazer uma melhor abrangência das respostas, isto se compararmos apenas os participantes que responderam a todos ou quase todos os tópicos (ludIT, GLNISTT, RAPPORTAGICO e RENOIR).

Como se vê claramente na tabela 6, e tal como já foi referido antes, os participantes humanos não responderam todos a todos os tópicos. A avaliação por cenários da mesma tabela evidencia que o número de tópicos em comum é baixo entre participantes, sobretudo entre os que responderam a um subconjunto dos 150 tópicos: os participantes Ângela Mota, que respondeu a 50 tópicos, e João Miranda, que respondeu a 40, partilham entre si apenas um décimo do total de tópicos, enquanto esses dois participantes com o participante Bruno Nascimento partilham somente 3 e 5 tópicos, respetivamente. Naturalmente, isso faz com que na avaliação por cenários (ver tabela 7), em termos de pontuação final, cada um desses participantes fique em terceiro lugar quando avaliado no seu cenário, não conseguindo mesmo assim superar os participantes ludIT e GLNISTT que deram em média um número maior de respostas por tópico do que esses participantes, e consequentemente um maior número de respostas corretas.

As tabelas 8 e 9 mostram, respetivamente, quantos participantes responderam ao mesmo tópico e quantos responderam corretamente ao mesmo tópico. Salienta-se que:

- não houve nenhum tópico que tenha sido respondido pelos sete participantes, apenas 16 foram respondidos por seis e houve pelo menos três participantes a responder a cada tópico, sendo que 18 foram respondidos apenas por três participantes (ver tabela 8);
- dos 16 tópicos respondidos por 6 participantes, apenas 3 foram respondidos corretamente também por 6 participantes, pouco mais de 20% dos tópicos foram respondidos corretamente por apenas 1 ou 2 participantes, e um dos tópicos não foi respondido corretamente por nenhum participante (ver tabela 9);
- existe apenas uma resposta dada pelos 6 participantes (que responderam ao mesmo tópico) e essa resposta está correta; a única resposta que foi dada por 5 participantes também está correta e mais de metade das respostas corretas foram dadas por um único participante, o que não quer dizer que tenha sido sempre o mesmo (ver tabela 10).

Cenário	Participante (Corrida)	$\|T\|$	$\|R\|$	$\|R\|/\|T\|$	$\|C\|$	$\|\tilde{C}\|$
Págico	ludIT	150	1387	9,25	1065	34
	GLNISTT	148	1016	6,86	661	52
	João Miranda	40	101	2,52	80	3
	Ângela Mota	50	157	3,14	88	3
	RAPPORTAGICO (3)	114	1730	15,18	208	13
	RAPPORTAGICO (2)	115	1736	15,1	203	13
	RAPPORTAGICO (1)	116	1718	14,81	181	11
	Bruno Nascimento	18	34	1,89	23	1
	RENOIR (1)	150	15000	100	436	38
	RENOIR (3)	150	15000	100	398	29
	RENOIR (2)	150	15000	100	329	25
GLNISTT	ludIT	148	1384	9,35	1063	34
	GLNISTT	148	1016	6,86	661	52
	João Miranda	39	100	2,56	79	3
	Ângela Mota	48	152	3,17	85	3
	RAPPORTAGICO (3)	112	1702	15,2	206	13
	RAPPORTAGICO (2)	113	1708	15,12	201	13
	RAPPORTAGICO (1)	114	1692	14,84	179	11
	Bruno Nascimento	18	34	1,89	23	1
	RENOIR (1)	148	14800	100	433	38
	RENOIR (3)	148	14800	100	395	29
	RENOIR (2)	148	14800	100	327	25
AM	ludIT	50	585	11,7	490	9
	GLNISTT	48	430	8,96	289	19
	Ângela Mota	50	157	3,14	88	3
	João Miranda	15	39	2,6	31	-
	RAPPORTAGICO (2)	44	743	16,89	105	8
	RAPPORTAGICO (3)	44	732	16,64	104	7
	RAPPORTAGICO (1)	44	722	16,41	85	7
	RENOIR (1)	50	4999	99,98	223	17
	RENOIR (3)	50	5000	100	194	13
	RENOIR (2)	50	5000	100	160	9
	Bruno Nascimento	3	5	1,67	3	-
JM	ludIT	40	430	10,75	344	10
	GLNISTT	39	342	8,77	224	18
	João Miranda	40	101	2,52	80	3
	RAPPORTAGICO (3)	30	488	16,27	59	5
	RAPPORTAGICO (2)	30	487	16,23	56	4
	Bruno Nascimento	5	12	2,4	8	-
	Ângela Mota	15	25	1,67	11	-
	RENOIR (1)	40	4002	100,05	128	16
	RAPPORTAGICO (1)	29	465	16,03	42	5
	RENOIR (3)	40	4000	100	122	13
	RENOIR (2)	40	4000	100	110	8
BN	ludIT	18	177	9,83	135	4
	GLNISTT	18	64	3,56	47	3
	Bruno Nascimento	18	34	1,89	23	1
	Ângela Mota	3	18	6	14	-
	João Miranda	5	15	3	11	-
	RAPPORTAGICO (3)	12	220	18,33	35	1
	RAPPORTAGICO (1)	12	179	14,92	28	1
	RAPPORTAGICO (2)	12	183	15,25	28	1
	RENOIR (1)	18	1800	100	60	1
	RENOIR (3)	18	1800	100	46	1
	RENOIR (2)	18	1800	100	29	1

Tabela 6: Estatísticas sobre as respostas.

Cenário	Participante (Corrida)	M	P	α	ϕ	$\tilde{P}$	O	K
Págico	ludIT	817,754	0,768	0,474	0,586	0,792	3442	3995,21
	GLNISTT	430,04	0,651	0,294	0,405	0,702	1767	2211,826
	João Miranda	63,366	0,792	0,036	0,068	0,822	202	287,139
	Ângela Mota	49,325	0,56	0,039	0,073	0,58	146	251,395
	RAPPORTAGICO (3)	25,008	0,12	0,092	0,104	0,128	29	297,003
	RAPPORTAGICO (2)	23,738	0,117	0,09	0,102	0,124	5	265,219
	RAPPORTAGICO (1)	19,069	0,105	0,08	0,091	0,112	22	224,72
	Bruno Nascimento	15,559	0,676	0,01	0,02	0,706	37	65,667
	RENOIR (1)	12,673	0,029	0,194	0,051	0,032	126	745,087
	RENOIR (3)	10,56	0,026	0,177	0,046	0,028	54	618,504
	RENOIR (2)	7,216	0,022	0,146	0,038	0,024	220	609,232
GLNISTT	ludIT	816,452	0,768	0,474	0,586	0,793	3438	3990,654
	GLNISTT	430,04	0,651	0,295	0,406	0,702	1767	2211,826
	João Miranda	62,41	0,79	0,035	0,067	0,82	202	286,583
	Ângela Mota	47,533	0,559	0,038	0,071	0,579	141	244,173
	RAPPORTAGICO (3)	24,933	0,121	0,092	0,104	0,129	29	295,614
	RAPPORTAGICO (2)	23,654	0,118	0,09	0,102	0,125	5	263,831
	RAPPORTAGICO (1)	18,937	0,106	0,08	0,091	0,112	22	223,331
	Bruno Nascimento	15,559	0,676	0,01	0,02	0,706	37	65,667
	RENOIR (1)	12,668	0,029	0,193	0,051	0,032	126	742,032
	RENOIR (3)	10,542	0,027	0,176	0,046	0,029	54	615,448
	RENOIR (2)	7,225	0,022	0,146	0,038	0,024	220	607,843
AM	ludIT	410,427	0,838	0,474	0,605	0,853	1868	2126,441
	GLNISTT	194,235	0,672	0,28	0,395	0,716	897	1091,941
	Ângela Mota	49,325	0,56	0,085	0,148	0,58	146	251,395
	João Miranda	24,641	0,795	0,03	0,058	0,795	90	125,472
	RAPPORTAGICO (2)	14,838	0,141	0,102	0,118	0,152	0	163,944
	RAPPORTAGICO (3)	14,776	0,142	0,101	0,118	0,152	11	174,111
	RAPPORTAGICO (1)	10,007	0,118	0,082	0,097	0,127	22	129,361
	RENOIR (1)	9,948	0,045	0,216	0,074	0,048	55	441,078
	RENOIR (3)	7,527	0,039	0,188	0,064	0,041	6	349,945
	RENOIR (2)	5,12	0,032	0,155	0,053	0,034	115	343,651
	Bruno Nascimento	1,8	0,6	0,003	0,006	0,6	6	9,667
JM	ludIT	275,2	0,8	0,436	0,564	0,823	1471	1616,256
	GLNISTT	146,714	0,655	0,284	0,396	0,708	725	882,214
	João Miranda	63,366	0,792	0,101	0,18	0,822	202	287,139
	RAPPORTAGICO (3)	7,133	0,121	0,075	0,092	0,131	0	111,136
	RAPPORTAGICO (2)	6,439	0,115	0,071	0,088	0,123	0	105,136
	Bruno Nascimento	5,333	0,667	0,01	0,02	0,667	12	23,25
	Ângela Mota	4,84	0,44	0,014	0,027	0,44	28	35,889
	RENOIR (1)	4,094	0,032	0,162	0,053	0,036	25	262,52
	RAPPORTAGICO (1)	3,793	0,09	0,053	0,067	0,101	12	69,886
	RENOIR (3)	3,721	0,03	0,155	0,051	0,034	15	242,387
	RENOIR (2)	3,025	0,028	0,139	0,046	0,03	112	261,187
BM	ludIT	102,966	0,763	0,5	0,604	0,785	410	504,441
	GLNISTT	34,516	0,734	0,174	0,281	0,781	110	152,762
	Bruno Nascimento	15,559	0,676	0,085	0,151	0,706	37	65,667
	Ângela Mota	10,889	0,778	0,052	0,097	0,778	48	62,667
	João Miranda	8,067	0,733	0,041	0,077	0,733	43	46,5
	RAPPORTAGICO (3)	5,568	0,159	0,13	0,143	0,164	16	54,438
	RAPPORTAGICO (1)	4,38	0,156	0,104	0,125	0,162	0	32,188
	RAPPORTAGICO (2)	4,284	0,153	0,104	0,124	0,158	5	35,522
	RENOIR (1)	2	0,033	0,222	0,058	0,034	47	126,855
	RENOIR (3)	1,176	0,026	0,17	0,044	0,026	0	69,105
	RENOIR (2)	0,467	0,016	0,107	0,028	0,017	36	69,857

Tabela 7: Avaliação dos participantes no Págico.

# Participantes	# Tópicos
6	16
5	59
4	57
3	18

Tabela 8: Participantes que responderam ao mesmo tópico.

# Participantes	# Tópicos
6	3
5	28
4	41
3	45
2	20
1	12

Tabela 9: Participantes que responderam corretamente ao mesmo tópico.

6 Comparação pessoa vs. máquina

Ao contrário do que seria desejável, como se realça no balanço do Págico (Santos et al., 2012), a participação no Págico não foi suficientemente grande para se poderem tirar conclusões comparativas fiáveis entre humanos e sistemas, ou mesmo entre sistemas.

Ainda assim, exploramos nesta seção alguns pontos de partida para uma análise futura mais profunda.

6.1 Há tópicos mais difíceis?

O facto de nem todos os participantes terem respondido aos mesmos tópicos dificulta a análise sobre se haveria tópicos mais difíceis do que outros, e se essa dificuldade é sensível ao tipo de participação. Uma primeira tentativa no sentido de aferir essa dificuldade é observando os tópicos com mais respostas corretas para cada um dos tipos de participação.

A tabela 11 mostra os cinco tópicos onde se verifica o maior número de respostas corretas dadas exclusivamente pelos participantes humanos, e em que tanto para as respostas enviadas como para as respostas corretas dos participantes foram dados: o total de respostas (T), o número de respostas dadas exclusivamente pelos participantes humanos (H), exclusivamente pelos sistemas (S) e dadas por ambos os tipos de participantes (HS). Esses, aliás, são também os tópicos com mais respostas dadas exclusivamente pelos participantes humanos, mas não pela mesma ordem: o tópico com mais respostas corretas exclusivamente humanas (**tópico 106** [Vice-reis da Índia Portuguesa]) é o terceiro tópico com mais respostas exclusivamente humanas.

# Participantes	# Respostas	# Corretas
6	1	1
5	1	1
4	42	24
3	126	59
2	792	115
1	31523	220

Tabela 10: Total de participantes que deu a mesma resposta.

Tal como mostra a tabela 12, dos tópicos que reúnem o maior número de respostas corretas exclusivamente dadas por participantes humanos, apenas um se encontra também entre os cinco que obtiveram mais respostas corretas dadas exclusivamente por sistemas (**tópico 19** [Tribos indígenas que vivem na Amazónia]). Ao contrário do que acontece com as respostas dadas exclusivamente por humanos, os cinco tópicos onde existe um maior número de respostas corretas dadas exclusivamente por sistemas não são os tópicos com maior número de respostas enviadas pelos sistemas.

O tópico 19 é também o único que está entre os tópicos que reuniram maior número de respostas corretas dadas por ambos os tipos de participante, sendo o tópico com mais respostas enviadas e também corretas nesse caso (ver tabela 13). Os tópicos com maior número de respostas corretas de ambos os tipos de participação são os mesmos que têm o maior de número respostas dadas por ambos os tipos de participante.

Uma vez que entre os tópicos com maior número de respostas corretas exclusivamente humanas, exclusivamente automáticas e de ambos os tipos de participação existe apenas um que é comum aos três casos, esse facto parece sugerir que existem tópicos para os quais participantes humanos encontram mais facilmente as respostas corretas, outros em que os sistemas serão mais bem sucedidos a encontrar as respostas corretas e, finalmente, ainda outros para os quais é indiferente se são humanos ou máquinas a tentar encontrar as respostas para eles. No futuro, analisar as diferenças entre estes tópicos, as suas respostas e as suas justificações, poderia ser um bom ponto de partida para identificar os tipo de tópicos onde é mais essencial melhorar os sistemas de modo a que estes possam auxiliaram mais os humanos no que precisam.

Além dos tópicos onde houve mais respostas corretas, observámos os tópicos sem respostas corretas para cada uma dos três casos acima referidos, uma vez que isso demonstra alguma

ID	Tópico	Enviadas				Corretas			
		T	*H*	*A*	*HA*	*T*	*H*	*A*	*HA*
106	Vice-reis da Índia Portuguesa	262	83	170	9	88	**78**	1	9
147	Museus em capitais de países lusófonos	285	86	198	1	65	**65**	0	0
144	Locais referidos n Os Lusíadas	351	85	265	1	62	**60**	1	1
19	Tribos indígenas que vivem na Amazônia.	250	59	160	31	115	**56**	35	24
16	Membros da igreja associados à Teologia da Libertação.	211	51	153	7	48	**37**	6	5

Tabela 11: Tópicos com mais respostas corretas exclusivamente de participações humanas

ID	Tópico	Enviadas				Corretas			
		T	*H*	*S*	*HS*	*T*	*H*	*S*	*HS*
135	Aves de Angola	154	12	141	1	54	10	**44**	0
19	Tribos indígenas que vivem na Amazônia.	250	59	160	31	115	56	**35**	24
90	Filmes brasileiros premiados na categoria Montagem.	211	14	190	7	34	8	**19**	7
13	Dinossauros carnívoros que habitaram o Brasil.	182	11	166	5	23	6	**12**	5
104	Pesquisadores do folclore brasileiro	203	14	179	10	33	13	**11**	9

Tabela 12: Tópicos com mais respostas corretas exclusivamente de sistemas

dificuldade por parte dos participantes em encontrar as respostas.

Existem dez tópicos sem respostas corretas dadas exclusivamente por participantes humanos, sendo que para um deles também não existem respostas corretas dadas por sistemas (**tópico 53** [Toureiros a cavalo de países lusófonos com carreira internacional]). Esses tópicos encontram-se na tabela 14. Como se pode ver na tabela, em três desses tópicos (**tópico 4** [Mulheres violoncelistas de língua portuguesa], **tópico 5** [Flautistas que se naturalizaram brasileiros ou portugueses] e **tópico 107** [Dioceses católicas de Moçambique]) não existem também respostas dadas apenas por participantes humanos; em quatro desses tópicos também não houve respostas corretas exclusivas de sistemas.

Cerca de um quarto dos tópicos (36) não tem respostas corretas dadas por sistemas (seja exclusivamente ou não) e 38% dos tópicos (57) tem respostas corretas dadas por um ou outro tipo de participação, mas não pelos dois.

De alguma forma, a ausência de respostas exclusivas de um dos tipos de participação demostra que esses tópicos são mais difíceis para esse tipo de participante, ou que pelo menos a dificuldade poderá ser semelhante ao do outro tipo de participação se as respostas forem comum aos dois.

6.2 Comparação por temas

No Págico os tópicos foram classificados em temas e grandes temas. Assim, mostramos nas tabelas 15 e 16 o desempenho comparativo entre sistemas e participantes humanos, em termos de pontuação final (M) e de precisão (P), discriminado por grande tema e tema, respetivamente.

Como é facilmente constatável, tanto participantes humanos como sistemas tiveram a melhor precisão no tópicos de geografia, se bem que a pior precisão dos primeiros foi em ciência e as dos segundos em política. No entanto, ao nível da pontuação final, os participantes humanos sairam-se significativamente melhor nos tópicos de letras, enquanto sistemas continuaram a ser melhores em geografia. Dado que o RENOIR é um sistema vocacionado para a recolha de informação geográfica, talvez não seja de espantar esse resultado.

Tema	*M*		*P*	
	Hum.	Auto.	Hum.	Auto.
Letras	590.72	5.24	71.52	1.90
Artes	324.80	4.48	71.07	2.46
Geografia	268.88	8.86	71.70	3.62
Cultura	205.34	2.19	67.11	2.05
Política	107.58	0.77	65.60	1.39
Desporto	104.31	1.14	63.22	1.75
Ciência	59.08	1.88	61.54	2.57
Economia	45.10	0.32	71.59	1.61

Tabela 15: Pontuação final (M) e precisão (P) por tema e tipo de participação.

6.3 Comparação por localização

A comparação entre sistemas e participantes relativamente à classificação geográfica dos tópicos mostra que a melhor pontuação final para

ID	Tópico	Enviadas				Corretas			
		T	*H*	*S*	*HS*	*T*	*H*	*S*	*HS*
19	Tribos indígenas que vivem na Amazônia.	250	59	160	31	115	56	35	**24**
62	Praias de Portugal boas para a prática de surf	161	7	134	20	30	5	6	**19**
7	Guitarristas portugueses que também foram compositores.	242	26	197	19	34	17	0	**17**
11	Filmes sobre o cangaço.	223	21	185	17	41	20	4	**17**
79	Povos indígenas brasileiros considerados extintos.	199	29	153	17	49	27	6	**16**

Tabela 13: Tópicos com mais respostas corretas de ambos os tipos de participação

ID	Tópico	Enviadas				Corretas			
		T	*H*	*S*	*HS*	*T*	*H*	*S*	*HS*
4	Mulheres violoncelistas de língua portuguesa	242	0	240	2	3	**0**	1	2
5	Flautistas que se naturalizaram brasileiros ou portugueses.	203	0	201	2	3	**0**	1	2
71	Doenças presentes no Brasil no século XVII	197	5	189	3	2	**0**	1	1
94	Parques nacionais de Moçambique	172	1	167	4	4	**0**	0	4
107	Dioceses católicas de Moçambique	185	0	178	7	7	**0**	1	6
108	Jogadores de futebol nascidos em Cabo Verde que representaram a seleção portuguesa	141	2	136	3	2	**0**	0	2
111	Padres católicos que estão ou estiveram ativos em Timor	164	1	158	5	4	**0**	1	3
121	Frutos de Angola	125	2	117	6	4	**0**	1	3
132	Deputados da FRELIMO	188	1	185	2	1	**0**	0	1
153	Toureiros a cavalo de países lusófonos com carreira internacional	229	4	225	0	0	**0**	0	0

Tabela 14: Tópicos sem respostas corretas exclusivamente humanas

Subtema	*M*		*P*	
	Hum.	Auto.	Hum.	Auto.
história	407,42	3,89	72,37	1,93
geografia	197,15	8,56	70,92	4,05
cinema	137,93	5,04	84,10	4,89
demografia	135,19	11,15	85,03	12,12
literatura	123,60	0,57	67,91	1,30
política	107,58	0,77	65,60	1,39
desporto	104,31	1,14	63,22	1,75
música	96,66	1,10	57,53	1,69
antro./folc.	76,33	1,32	62,56	2,36
religião	70,92	1,08	70,92	3,08
cultura	50,24	0,02	67,89	0,62
televisão	49,42	0,09	91,53	0,98
artes	47,87	0,00	72,53	0,00
economia	45,10	0,32	71,59	1,61
filosofia	34,52	0,33	73,44	2,99
linguística	33,78	0,46	66,23	1,92
culinária	30,22	0,46	67,16	2,20
arquit./urb.	25,78	0,37	66,10	1,60
zoologia	19,70	7,50	72,97	12,30
jornalismo	18,23	0,04	67,50	0,78
ciência	13,33	0,00	66,67	0,00
saúde	10,62	0,04	55,88	0,70
geologia	8,76	0,00	48,65	0,00
ensino	6,05	0,00	55,00	0,26
botânica	4,92	0,07	61,54	1,32
artes plásticas	3,85	0,35	38,46	2,93
matemática	3,20	0,00	80,00	0,43

Tabela 16: Pontuação final (M) e precisão (P) por subtema e tipo de participação.

ambos os tipos de participação foi nos tópicos sobre o Brasil (veja-se a tabela 17). Isso talvez se deva ao facto de a maioria dos tópicos estar classificado com esse local, e que, como tal, terá à partida um maior número de respostas associado.

Ao nível da precisão, os participantes humanos obtiveram o melhor desempenho nos tópicos sobre a Guiné-Bissau, enquanto os sistemas obtiveram um melhor resultado nos temas de Angola.

7 Comentários finais

Com o Págico foram dados os primeiros passos no sentido de comparar o desempenho de humanos e sistemas numa tarefa de pesquisa de informação na Wikipédia. Embora o objectivo não tenha sido criar uma competição entre humanos e sistemas, mas sim uma colaboração entre ambos a fim de no futuro criar melhores sistemas que possam ajudar os humanos nessa tarefa, apresentámos neste artigo resultados detalhados, mas ainda superficiais, sobre a participação no Págico.

Além de ser necessário no futuro olhar mais aprofundadamente para estes resultados e de analisar as participações de outras perspetivas, realçamos aqui alguns pontos que talvez valha a

Lugar	*M*		*P*	
	Hum.	Auto.	Hum.	Auto.
Brasil	462.28	9.73	72.69	3.08
Lusofonia	275.89	1.47	61.86	1.22
Portugal	202.75	2.50	73.73	2.75
Geral	64.46	0.10	65.77	0.87
Moçambique	36.91	0.29	68.35	1.22
Angola	36.05	3.87	69.33	5.23
Macau	23.44	0.42	75.61	2.44
Cabo Verde	19.88	0.19	76.47	1.38
Timor	13.83	0.83	62.86	4.17
Guiné Bissau	5.44	0.00	77.78	0.39
São Tomé e Príncipe	4.45	0.03	63.64	1.14

Tabela 17: Pontuação final (M) e precisão (P) por localização e tipo de participação.

pena explorar:

- caracterizar os tópicos com mais e menos respostas corretas para cada tipo de participação - em (Simões, Costa e Mota, 2012) é feita uma caracterzição pelo número de palavras e de documentos sem ter em conta o tipo de participação;
- apresentar estatísticas de participação humana: quanto tempo e qual a ordem pela qual os participantes humanos tentaram responder, se alteraram a ordem pré-estabelecida, se tentaram responder primeiro a tópicos de um determinado tema e só depois passar a outro, etc.. Este trabalho, em parte foi iniciado em (Costa, Mota e Santos, 2012);
- avaliar as medidas de avaliação para ver até que ponto são realmente úteis para julgar a qualidade das respostas dos participantes.

Para tal, os interessados em estudar estas questões poderão consultar resultados adicionais disponibilizados no sítio do Págico, bem como usar o pacote do Págico, o Cartola, descrito em (Simões, Costa e Mota, 2012) e o SIGA para obter ainda mais resultados.

Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1,3/C/NAC, pela UMIC e pela FCCN, e em 2011 pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN).

Agradeço à restante organização do Págico pelas várias sugestões de medidas de avaliação, bem como pelas discussões sobre as mesmas e os demais aspetos relacionados com a avaliação dos participantes.

Estou também agradecida aos revisores convidados, Luísa Coheur e Paulo Gomes, pelas suas críticas construtivas que ajudaram a melhor significativamente o artigo.

Referências

Cardoso, Nuno. 2012. Medindo o precipício semântico. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Coheur, Luísa e Ângela Costa. 2012. Do tópico às respostas: do processo humano à sua simulação. *Linguamática*, 4(1), Abril, 2012. Neste volume.

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.

Costa, Luís, Paulo Rocha, e Diana Santos. 2007. Organização e resultados morfolímpicos. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, capítulo 2, pp. 15–33.

Freitas, Cláudia. 2012. A lusofonia na wikipédia em 150 tópicos. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Freitas, Cláudia, Paulo Rocha, Cristina Mota, Luís Costa, e Diana Santos. 2012. O que é uma resposta? Notas de uns avaliadores estafados. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Gonçalo Oliveira, Hugo, Cristina Mota, Cláudia Freitas, Diana Santos, e Paula Carvalho. 2008. Avaliação à medida no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 97–129, 31 de Dezembro, 2008.

Miranda, João. 2012. Desafios na recolha de informação baseada na Wikipédia portuguesa com o Págico. *Linguamática*, 4(1), Abril, 2012. "Neste volume".

Mota, Cristina, Alberto Simões, Cláudia Freitas, Luís Costa, e Diana Santos. 2012. Págico: Evaluating Wikipedia-based information retrieval in Portuguese. Em *Language Resources and Evaluation Conference*, Maio, 2012.

Oliveira, Hugo Gonçalo, Cristina Mota, Cláudia Freitas, Diana Santos, e Paula Carvalho. 2008. Avaliação à medida no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 97–129, 31 de Dezembro, 2008.

Rodrigues, Ricardo, Hugo Gonçalo Oliveira, e Paulo Gomes. 2012. Uma abordagem ao Págico baseada no processamento e análise de sintagmas dos tópicos. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Santos, Diana. 2012. Porquê o Págico? Razões para uma avaliação conjunta. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Santos, Diana, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Julia Schulz, Yvonne Skalban, e Erik Tjong Kim Sang. 2010. GikiCLEF: Crosscultural Issues in Multilingual Information Access. Em Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, Maio, 2010. European Language Resources Association (ELRA).

Santos, Diana, Cristina Mota, Alberto Simões, Luís Costa, e Cláudia Freitas. 2012. Balanço do Págico e perspetivas de futuro. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Simões, Alberto, Luís Costa, e Cristina Mota. 2012. Tirando o chapéu à Wikipédia: A coleção do Págico e o Cartola. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Veiga, Arlindo, Carla Lopes, Dirce Celorico, Jorge Proença, Fernando Perdigão, e Sara Candeias. 2012. O desafio da participação humana do IT-Coimbra no Págico. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Balanço do Págico e perspetivas de futuro

Diana Santos
Linguateca/FCCN &
Universidade de Oslo
d.s.m.santos@ilos.uio.no

Cristina Mota
Linguateca/FCCN
cmota@ist.utl.pt

Alberto Simões
Instituto de Letras e Ciências Humanas
Universidade do Minho
ambs@ilch.uminho.pt

Luís Costa
Linguateca/FCCN
luis.f.kosta@gmail.com

Cláudia Freitas
Linguateca/FCCN & PUC-Rio
maclaudia.freitas@gmail.com

Resumo

Uma avaliação só está concluída quando se faz um balanço e se tiram ilações para o futuro. Neste artigo discutimos o que foi obtido com o Págico, o que podia ter sido melhor, e propomos o que nos parece interessante de realizar numa próxima edição. Tentámos estruturar o balanço segundo as vertentes descritas na motivação do Págico, resumindo e opinando sobre outros artigos nesta edição, e depois fazendo uma apreciação crítica da participação e dos resultados. Descrevemos ainda formas de utilizar, numa fase pós-Págico, o extenso trabalho desenvolvido, quer na avaliação de novos sistemas, quer no ensino e/ou divulgação de temas de cultura lusófona por exemplo no estrangeiro. Em relação à área de usabilidade, deixamos em aberto o desafio de como melhorar significativamente a interface do SIGA de forma a poder ser usada em casos reais.

Palavras chave

Recolha de informação, Resposta a perguntas, Avaliação, Cooperação pessoa-máquina, Wikipédia, Usabilidade

1 Apresentação

Como primeiro acontecimento dedicado a avaliar e estudar a procura de informação na wikipédia em português, o Págico terá merecido o seu lugar na história, mas a participação ficou muito aquém das expetativas, sobretudo se tivermos em conta que se dedicava a todo o mundo lusófono: participaram três grupos de investigação que fazem investigação em recolha de informação (RI) e resposta a perguntas (RAP), mas um deles sem sistema e como equipa humana, e ainda outras quatro equipas humanas, das quais apenas uma respondeu a todos os tópicos e as restantes não responderam a mais de um terço dos tópicos cada uma.

De facto, a participação é comparável ao que por exemplo é relatado na tese de Rachel Aires Aires (2005), em que seis utilizadores criaram corpora personalizados para o estudo em questão, ou, em geral, estudos feitos em contextos de doutoramento, e muito inferior, a nível de participação de grupos de investigação, ao que foi o caso nas anteriores avaliações conjuntas organizadas pela Linguateca, como as Morfolimpíadas (Costa, Rocha e Santos, 2007), com sete sistemas participantes, ou o HAREM (Santos e Cardoso, 2007; Mota e Santos, 2008), em que participaram nove sistemas.

Podemos tentar entender esse decréscimo de duas maneiras, não necessariamente excludentes. Por um lado, temos consciência de que a tarefa proposta no Págico é muito mais complexa que a das avaliações conjuntas anteriores organizadas pela Linguateca, o que pode ter assustado alguns. A nossa intenção foi realmente propor um desafio que fosse o mais próximo possível dos "desafios do mundo real," em oposição às tarefas da Morfolimpíadas ou do HAREM (veja-se a secção 8 sobre a comparação com avaliações conjuntas anteriores). Por outro lado, podemos também supor que, aparentemente, as áreas de recolha de informação e de resposta a perguntas em Portugal e no Brasil não consideraram o desafio válido ou interessante, nem as áreas de estudos de usabilidade ou de estudos de utilizadores se interessaram pela nossa iniciativa. A área de PLN, por outro lado, achou provavelmente esta uma tarefa difícil demais, aliás como aconteceu no GikiCLEF e no GikiP, onde a participação de língua portuguesa foi mínima. Isto pode significar ainda que estas disciplinas não estão focadas no português, ou que talvez devêssemos ter uma base de organizadores muito maior para que cada grupo desafiasse os seus membros, e para que dessem

o aval científico nas áreas respetivas (RI, RAP e usabilidade).

Entre outras coisas, este artigo pretende de certa forma contrariar essa segunda visão, e tentar que, à posteriori, o trabalho feito possa de facto ser usado nessas áreas e por esses investigadores (veja-se a secção 7). Mas antes disso queremos olhar para o que foi feito e ver o que aprendemos e o que podemos apresentar para que outros possam aprender.

2 A wikipédia melhorou?

Certamente que não! Ou seja, por muito que tenhamos estudado e interagido com esta enciclopédia comum, de forma alguma isso teve impacto na realidade.[1] Mas é nossa esperança que, ao distribuirmos e iniciarmos um estudo quantitativo de várias questões, veja-se (Simões, Costa e Mota, 2012), possamos entusiasmar outros a fazerem algo semelhante ou muito mais interessante.

Outra das coisas que será possível fazer, se houver interesse, é medir, de acordo com critérios semelhantes ou pelo menos comparáveis, versões futuras da Wikipédia, lançando, quem sabe, um wikiavaliómetro, à semelhança do barómetro das línguas românicas da União Latina, referido por Santos (2012), mas referindo-se ao português e à wikipédia em português mais especificamente.

3 A comparação pessoa-máquina

O Págico acabou por não dar um contributo especial nesta matéria—embora tenhamos progredido no desenvolvimento de ferramentas que podem levar a esse objetivo, ao melhorar o SIGA e equipá-lo com capacidades de reflexão ou investigação do comportamento dos utilizadores, como ilustrado já em (Costa, Mota e Santos, 2012).

A principal razão foi a já mencionada falta de participantes, sobretudo automáticos, mas também humanos, que não nos permite generalizar com um mínimo de confiança, e, por outro lado, a confirmação de que os seres humanos ainda não têm par na resposta ao tipo de perguntas do Págico. Contudo, houve algumas respostas corretas encontradas pelos sistemas e não propostas pela participação humana, o que leva a esperar que de facto o concurso, no sentido de ajuda, já é e será cada vez mais benéfico na resposta a perguntas a uma base de grande informação.

Foi de qualquer forma importante que, mais uma vez, todos fôssemos obrigados a refletir na questão extremamente complexa do que é uma resposta (Freitas et al., 2012) e na dificuldade de delimitação rigorosa do que é útil ou apropriado na procura ou descoberta de informação sobre um dado tema.

4 O realismo da tarefa

Uma das questões metodologicamente mais complexas na organização de uma avaliação conjunta é a obtenção de uma tarefa finita e bem delimitada que seja por outro lado passível de repetição e extensão. Que não seja simplesmente uma demo ou uma curiosidade, mas que seja de certa forma representativa de problemas práticos e autênticos na vida de utilizadores da wikipédia (como fonte de informação sobre a cultura lusófona).

A primeira coisa que tivemos de decidir e que, de certa forma, contraria o realismo da tarefa, mas que era essencial no nosso caso, foi a limitação à versão em português. Todos nós sabemos da possibilidade de navegação entre várias línguas na wikipédia, portanto não é realista obrigar a procurar só em português. Mas tínhamos razões de sobra para fazer esta escolha: por um lado, criar uma coleção da wikipédia com todas as línguas potenciais dos participantes incluindo o inglês seria uma tarefa demasiado grande para os nossos meios; por outro—e esta talvez seja a razão mais importante—, fizemo-lo no GikiCLEF e acabámos por não conseguir medir o impacto ou interesse da parte portuguesa.

Já referimos em diversos outros artigos (por exemplo, em (Santos, 2012) e (Mota et al., 2012)), mas importa aqui novamente realçar, que também o desenvolvimento de um sistema de navegação na wikipédia para indicar respostas e justificações, que concorresse com a forma humana e habitual com que os participantes contactam e interagem com esse recurso, não foi fácil, e possivelmente nem mesmo bem sucedido.

Este é um dado que precisamos de levar em conta mais tarde, se viermos a organizar mais avaliações conjuntas com participação humana. Se, para sistemas automáticos, é só definir uma sintaxe rigososa de entrada e saída, e escrever validadores que a verifiquem ou corrijam, a situação é totalmente diferente quando queremos que participantes humanos não sejam impedidos ou contrariados, em vez de ajudados, numa dada

[1] No entanto, foram pontualmente feitas, e marcadas para fazer no futuro, correções ortográficas, gramaticais, de conteúdo e a hiper-ligações com problemas diversos.

tarefa. Deveriam ter sido feitos estudos de usabilidade e ter sido dada muita mais atenção à forma de desenvolver sistemas realmente apropriados à tarefa em mente, e o facto de termos (todos, organizadores e participantes) tido pouco tempo e termos dado pouco retorno acabou por espantar muitos participantes, ou levar a que a maior parte deles usasse, não o nosso sistema, mas a interface normal da wikipédia. É interessante ver que isto aponta para duas linhas de desenvolvimento que já têm sido mencionadas:

- tentar diminuir a novidade ou diferença nas interfaces: idealmente, apenas adicionar algo àquilo que os utilizadores já conhecem e de que gostam, e não obrigá-los a criar novos hábitos ou raciocínios;
- não pedir para fazer mais do que é preciso... veja-se a interface do Webpaper descrita em (Simões, Rocha e Fonseca, 2009), que aliás reuniu muito mais participantes que o Págico, mostrando que atividades lúdicas, ou didacto-lúdicas, têm um grande potencial para recolher informação sobre participantes humanos.

Uma outra questão que surgiu, em particular através da chamada de atenção da Belinda Maia, foi a de os temas escolhidos serem maioritariamente de letras, e não de ciência, economia ou técnica/tecnologia.

A questão aqui é até que ponto existe ciência em português, ou seja, até que ponto a informação que tentaríamos obter era escrita de raiz em português ou especialmente relacionada à cultura lusófona.

Mas esta é uma área e atividade que, a nível pedagógico nos próprios países lusófonos, poderá ter um impacto fundamental e mais importante do que aquele relacionado com a cultura lusófona para estrangeiros, e que fica pois agendada como ideia para o futuro.

5 Os problemas do reuso

É evidente que sempre que não se começa do princípio, mas se usa algo já desenvolvido, isso tem vantagens. Mas é preciso também referir que nem tudo são rosas numa tal abordagem, sobretudo se os novos desenvolvedores não são os mesmos do sistema anterior, como foi o caso do SIGA.

Assim, a escolha da forma de desenvolvimento de um dado sistema passa a obedecer a dois princípios que por vezes são contraditórios:

1. minimizar as alterações ao que já está feito, procedendo de forma incremental;
2. adicionar novas funcionalidades de acordo com o mais adequado ao utilizador.

Esta questão, que não é de fácil resolução, teve impacto nas três mudanças principais realizadas ao SIGA: a adição de utilizadores humanos, já comentada acima, a melhoria da interface de avaliação, e a apresentação dos resultados com novas medidas.

Nos três casos poderíamos ter desenvolvido soluções mais inovadoras e capazes. Todavia, talvez não tivéssemos ainda realizado a própria avaliação, que foi aquela que foi conseguida num prazo tão curto.

6 Autocrítica

Existe uma série de pontos em que fizemos as opções erradas ou não conseguimos dar conta do recado, e que parece mais natural indicar aqui por atacado numa lista, sem tentar justificar ou desculpar. Obviamente, esses casos são automaticamente casos a melhorar, se houver um próximo Págico:

- Não houve qualquer sugestão de perguntas ou tópicos por parte dos participantes. Se tivéssemos conseguido que os tópicos fossem/fizessem parte de uma partilha de vários investigadores sobre questões que lhes interessavam e sobre as quais quereriam saber mais, e ao mesmo tempo na área ou em questões sobre as quais os seus sistemas brilhariam ou estavam especialmente interessados em ser avaliados, o processo teria sido muito melhor e tido muito mais participação.
- Não publicámos as medidas de avaliação a tempo de serem discutidas, internalizadas ou sequer tomadas em conta no desenvolvimento dos sistemas. De facto, foram publicadas só depois de os participantes terem enviado as suas respostas.
- A coleção do Págico foi disponibilizada em várias versões, e nem a última estava imume a problemas, conforme descrito em (Simões, Costa e Mota, 2012).
- Devido a vários problemas detetados demasiado tarde, a interface de participação humana foi alterada várias vezes durante o próprio mês em que o Págico esteve aberto, o que pode ter levado a confundir os utilizadores, e de dificultar os nossos estudos da sua interação.

- Ao contrário das nossas intenções, não conseguimos publicar um manual de utilização do sistema para os participantes humanos, o que, estamos convencidos, afastou alguns inscritos e muitos que poderiam ter tentado se fossem mais ajudados.
- Devido a um problema apenas descoberto tarde, algumas respostas não foram sequer avaliadas, resultando em erros nos resultados finais.
- Muitas das modificações e melhorias feitas ao sistema de avaliação foram-no à posteriori, não tendo os avaliadores a possibilidade de delas beneficiar.
- Devido à enorme quantidade de respostas, não foi possível usar uma das funcionalidades mais interessantes do SIGA, nomeadamente a avaliação sobreposta e a análise subsequente de possíveis conflitos, a não ser num número muito diminuto de casos, como referido em (Freitas et al., 2012)
- Devido a um engano, parte da avaliação sobreposta foi feita conhecendo a anterior avaliação: ou seja, as respostas a avaliar foram atribuidas referindo que "estas são as respostas duvidosas do avaliador Y", o que impediu observar se também levantariam dúvidas a outros avaliadores não precavidos.
- Por causa dos prazos, foi necessário publicar os resultados sem fazer uma revisão completa às respostas, o que implicou que estamos conscientes de ainda haver erros no material disponibilizado.
- Não foi possível traduzir toda a interface do SIGA, que estava em inglês, para português, nem vice-versa no que se refere às funcionalidades novas, que estão apenas em português, nem documentar exaustivamente as ditas.
- Também não foi possível incluir ou processar convenientemente as corridas não oficiais dos sistemas automáticos, o que claramente melhoraria o Cartola (Mota, 2012; Simões, Costa e Mota, 2012).

Por todas estas imprecisões ou faltas, estamos convencidos de que seria muito interessante ter um período de consolidação pós-Págico em que tanto os recursos, como o sistema, como a coleção pudessem ser polidos, melhorados e investigados em mais detalhe—não só por nós, mas por todos quantos acham o assunto interessante. Consideramos que os problemas e imperfeições existentes no recurso são em muito superados pelo facto de não guardarmos o nosso trabalho só para nós ou esperarmos que esteja perfeito para disponibilizar... pelo contrário, disponibilizamo-lo a todos assim que o consideramos útil (o que não nos impede de continuar a melhorar e criar novas versões), para que ajudem a melhorá-lo e possam aprender com os nossos erros também.

7 Contributos para o futuro

Como já dissemos, o Págico foi organizado e concluído num tempo recorde, o que faz com que muita prospeção sobre os dados recolhidos, que nós teríamos gostado de fazer, ficou para o futuro.

Mas além disso gostávamos de mencionar aqui algumas ideias de aproveitar criativamente o material, em ocasiões posteriores:

- refazer a coleção noutras datas, e confirmar / reclassificar as respostas encontradas nessa altura, muito provavelmente pedindo a novas equipas/pessoas para encontrar as respostas;
- com base em temas ou super-temas, tentar criar perguntas automaticamente, à semelhança da avaliação conjunta QG ("question generation") (Rus et al., 2012), veja-se `http://www.questiongeneration.org/`;
- usar o SIGA para adicionar mais tópicos e respostas de forma a ir criando uma base maior de perguntas respondidas pela wikipédia;
- usando as mesmas perguntas e respostas, mas com novas pessoas, verificar como é que elas reagiriam para justificar ou negar uma dada resposta: ou seja, criar algo parecido com a AVE (Rodrigo, Peñas e Verdejo, 2009), para identificar processos típicos ou comuns de raciocínio humano, e também clarificar a dificuldade ou não de diferentes pares de perguntas e respostas;
- desenvolver sistemas interativos que apresentassem, para cada tópico, o conjunto de respostas de maneira agregada e satisfatória.

Todas estas iniciativas poderiam aumentar o valor do Cartola e capitalizar o trabalho feito no Págico, e oferecemo-las a quem as quiser desenvolver.

8 Comparação com avaliações conjuntas anteriores

Finalmente, parece-nos apropriado fazer alguns comentários baseados na longa experiência de

avaliações conjuntas que a Linguateca iniciou há precisamente dez anos.

Não há dúvida que essa organização é trabalhosa e tem de ser independente dos participantes; por outro lado, desde o início que quanto mais associados ou ligados à Linguateca (ou à organização), maior a probabilidade de um grupo ou investigador participar.

Se por um lado é natural que grupos próximos partilhem opiniões científicas sobre o progresso na área, o perigo da participação próxima é que isso pode levar a que as avaliações conjuntas sejam vistas como apenas valendo a pena para a própria Linguateca e para fomentar os nossos objetivos, desvirtuando pois a vertente de serviço à comunidade, que é o que nos move.

Em 2008, fizemos uma consulta à comunidade, e a resposta que tivemos foi a de que alguns grupos (dois) estariam interessados em avaliar sistemas de deteção de terminologia. Contudo, nós não considerámos apropriado fazer uma avaliação conjunta nesse domínio visto que não havia nenhuma forma independente de obter recursos dourados, e pareceu-nos que uma avaliação à posteriori iria ser demasiado subjetiva. Além disso, e ao contrário do Págico, não conseguiríamos juntar resultados que fossem relevantes para futuras edições, ou seja, seria algo que apenas serviria para um dado conjunto de textos e domínios, fixo na altura.

Mas por outro lado estamos conscientes de que, e dado o decréscimo significativo do número de participantes das Morfolimpíadas para o HAREM para o CLEF/GikiCLEF e agora para o Págico, sem garantir uma base real de participantes suficientemente alargada não faz sentido a Linguateca organizar mais avaliações conjuntas.

Uma outra vertente que faz sentido comentar é o tamanho das coleções disponibilizadas e que se esperava que os sistemas processassem. De um conjunto de pequenos textos nas Morfolimpíadas para coleções de textos pequenas no HAREM passámos a coleções jornalísticas e depois para versões da wikipédia cada vez maiores. Se por um lado isso espelha o progresso no PLN, por outro pode também consistir um problema para arranjar participantes—como mencionado em (Cardoso, 2012), o sistema não conseguiu processar a coleçao toda em tempo útil.

Finalmente, repare-se que as tarefas em que tentamos avaliar um sistema são cada vez mais próximas das tarefas de um utilizador humano: analisar corretamente uma dada palavra fora do contexto em todas as interpretações morfológicas é algo que só se faz (?) na escola, estando claramente na província de sub-sistemas transparentes ao utilizador, para usar a terminologia de Gaizauskas (1998), classificar/identificar um nome próprio como mencionando uma pessoa, uma instituição, uma abstração ou um local é já algo que um ser humano faz "automaticamente", sem pensar, enquanto procurar informação sobre um tema para responder a perguntas concretas usando a wikipédia é algo que se faz conscientemente no mundo atual.

Dito isto podemos portanto identificar uma diminuição no número de participantes e de movimento à volta da avaliação – que, pensamos, congregou praticamente toda a comunidade no caso das Morfolimpíadas, pelo menos se virmos a audiência do encontro satélite do PROPOR 2003 na altura—e que no Págico foi mínima, e por outro lado um aumento na dificuldade da tarefa oferecida, tanto a nível de tamanho de recurso como a nível de comparação com desempenho humano.

9 Comentários finais

Pensamos que mesmo com as restrições temporais e de financiamento que a Linguateca sofreu, e que podemos adjetivar de drásticas sem exagero[2], o Págico conseguiu alguns resultados interessantes, e congregar mais alguns interessados na área da inter-relação entre a recolha de informação e o processamento computacional das línguas.

Pensamos que o grande contributo, além do início de uma avaliação científica da wikipédia em português, foi a criação do Cartola, que permitirá que muitos outros investigadores, no futuro, possam treinar os seus sistemas e/ou avaliá-los com base no material por nós coligido, além de fazer prospeção de outros assuntos que nós nem sequer tenhamos (ainda) abordado.

Também apelamos à criação de novas iniciativas, por exemplo com valor pedagógico, ou noutra áreas mais relacionadas com interesses específicos e em que a wikipédia pode ser ou vir a ser uma fonte importante, que possam (re)usar o nosso trabalho, e os sistemas desenvolvidos.

Agradecimentos

O trabalho aqui descrito enquadra-se no âmbito da Linguateca, co-financiada desde o seu início pelo Governo Português, pela União Europeia (FEDER e FSE), sob o contrato

[2]Em relação a 2011, o panorama está descrito em Santos (2011), em relação a 2012, não obtivemos qualquer financiamento.

POSC/339/1.3/C/NAC, pela UMIC e pela FCCN, e durante 2011, pela Fundação da Ciência e da Tecnologia (FCT) e pela Fundação para a Computação Científica Nacional (FCCN).

A Universidade de Oslo também contribuiu para a organização do Págico, e especificamente financiou substancialmente a organização do encontro do Págico. Um dos membros da organização foi parcialmente suportado pela bolsa da Fundação da Ciência e da Tecnologia SFRH/BPD/73011/2010. Também agradecemos à PUC-Rio e à universidade de Coimbra pelo apoio prestado.

Agradecemos sinceramente a todos os participantes, sem os quais o Págico teria sido inútil, ao Fernando Ribeiro e ao Hernâni Costa, da equipa da Linguateca, pelo retorno sobre o SIGA no que se refere à participação humana, e ao Paulo Rocha pelo esforço titânico na avaliação e pela cultura geral e imaginação empregada na escolha de tópicos.

Agradecemos também à comissão científica do presente volume, em particular ao Xavier Guinovart e ao António Teixeira, que, mais uma vez submetida a prazos sobre-humanos, conseguiu mesmo assim contribuir significativamente para a sua qualidade e para a variedade de ideias interessantes que podemos apresentar.

Referências

Aires, Rachel Virgínia Xavier. 2005. *Uso de marcadores estilísticos para a busca na Web em português.* Tese de doutoramento, ICMC - USP - São Carlos, Agosto, 2005. `http://www.linguateca.pt/documentos/TeseDoutRachelAires.pdf`.

Cardoso, Nuno. 2012. Medindo o precipício semântico. *Linguamática*, 4(1), Abril, 2012. Neste volume.

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.

Costa, Luís, Paulo Rocha, e Diana Santos. 2007. Organização e resultados morfolímpicos. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, capítulo 2, pp. 15–33.

Freitas, Cláudia, Paulo Rocha, Cristina Mota, Luís Costa, e Diana Santos. 2012. O que é uma resposta? Notas de uns avaliadores estafados. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Gaizauskas, Robert. 1998. Evaluation in language and speech technology. *Computer Speech and Language*, 12(4):249–62.

Mota, Cristina. 2012. Resultados págicos: participação, medidas e pontuação. *Linguamática*, 4(1), Abril, 2012. Neste volume.

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

Mota, Cristina, Alberto Simões, Cláudia Freitas, Luís Costa, e Diana Santos. 2012. Págico: Evaluating Wikipedia-based information retrieval in Portuguese. Em *Language Resources and Evaluation Conference*, Maio, 2012.

Rodrigo, Álvaro, Anselmo Peñas, e Felisa Verdejo. 2009. Overview of the answer validation exercise 2008. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, pp. 296–313, Berlim/Heidelberg. Springer.

Rus, Vasile, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, e Cristian Moldovan. 2012. A Detailed Account of The First Question Generation Shared Task Evaluation Challenge. *Dialogue & Discourse*, 3(2):177–204.

Santos, Diana. 2011. Relatório da Linguateca relativo ao ano de 2011. Relatório técnico, Linguateca/FCCN. `http://www.linguateca.pt/documentos/Relatorio2011LinguatecaFinal.pdf`.

Santos, Diana. 2012. Porquê o Págico? Razões para uma avaliação conjunta. *Linguamática*, 4(1), Abril, 2012. Neste volume.

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, Linguateca, 12 de Novembro, 2007.

Simões, Alberto, Luís Costa, e Cristina Mota. 2012. Tirando o chapéu à Wikipédia: A coleção do Págico e o Cartola. *Linguamática*, 4(1), Abril, 2012. Neste volume.

Simões, Alberto, Paulo Rocha, e Rúben Fonseca. 2009. Webpaper — más perguntas e boas respostas: a arte de interrogar. Em Paulo Dias, António José Osório, e Altina Ramos, editores, *O digital e o currículo*. Centro de Competência da Universidade do Minho, pp. 227–238, Maio, 2009.

Chamada de Artigos

A revista Linguamática pretende colmatar uma lacuna na comunidade de processamento de linguagem natural para as línguas ibéricas. Deste modo, serão publicados artigos que visem o processamento de alguma destas línguas.

A Linguamática é uma revista completamente aberta. Os artigos serão publicados de forma electrónica e disponibilizados abertamente para toda a comunidade científica sob licença *Creative Commons*.

Tópicos de interesse:

- Morfologia, sintaxe e semântica computacional
- Tradução automática e ferramentas de auxílio à tradução
- Terminologia e lexicografia computacional
- Síntese e reconhecimento de fala
- Recolha de informação
- Resposta automática a perguntas
- Linguística com corpora
- Bibliotecas digitais
- Avaliação de sistemas de processamento de linguagem natural
- Ferramentas e recursos públicos ou partilháveis
- Serviços linguísticos na rede
- Ontologias e representação do conhecimento
- Métodos estatísticos aplicados à língua
- Ferramentas de apoio ao ensino das línguas

Os artigos devem ser enviados em PDF através do sistema electrónico da revista. Embora o número de páginas dos artigos seja flexível sugere-se que não excedam 20 páginas. Os artigos devem ser devidamente identificados. Do mesmo modo, os comentários dos membros do comité científico serão devidamente assinados.

Em relação à língua usada para a escrita do artigo, sugere-se o uso de português, galego, castelhano, basco ou catalão.

Os artigos devem seguir o formato gráfico da revista. Existem modelos LaTeX, Microsoft Word e OpenOffice.org na página da Linguamática.

Datas Importantes

- Envio de artigos até: 15 de Abril de 2012
- Resultados da selecção até: 15 de Maio de 2012
- Versão final até: 31 de Maio de 2012
- Publicação da revista: Junho de 2012

Qualquer questão deve ser endereçada a: `editores@linguamatica.com`

Petición de Artigos

A revista Linguamática pretende cubrir unha lagoa na comunidade de procesamento de linguaxe natural para as linguas ibéricas. Deste xeito, han ser publicados artigos que traten o procesamento de calquera destas linguas.

Linguamática é unha revista completamente aberta. Os artigos publicaranse de forma electrónica e estarán ao libre dispor de toda a comunidade científica con licenza *Creative Commons*.

Temas de interese:

- Morfoloxía, sintaxe e semántica computacional
- Tradución automática e ferramentas de axuda á tradución
- Terminoloxía e lexicografía computacional
- Síntese e recoñecemento de fala
- Extracción de información
- Resposta automática a preguntas
- Lingüística de corpus
- Bibliotecas dixitais
- Avaliación de sistemas de procesamento de linguaxe natural
- Ferramentas e recursos públicos ou cooperativos
- Servizos lingüísticos na rede
- Ontoloxías e representación do coñecemento
- Métodos estatísticos aplicados á lingua
- Ferramentas de apoio ao ensino das linguas

Os artigos deben de enviarse en PDF mediante o sistema electrónico da revista. Aínda que o número de páxinas dos artigos sexa flexíbel suxírese que non excedan as 20 páxinas. Os artigos teñen que identificarse debidamente. Do mesmo modo, os comentarios dos membros do comité científico serán debidamente asinados.

En relación á lingua usada para a escrita do artigo, suxírese o uso de portugués, galego, castelán, éuscaro ou catalán.

Os artigos teñen que seguir o formato gráfico da revista. Existen modelos LaTeX, Microsoft Word e OpenOffice.org na páxina de Linguamática.

Datas Importantes

- Envío de artigos até: 15 de abril de 2012
- Resultados da selección: 15 de maio de 2012
- Versión final: 31 de maio de 2012
- Publicación da revista: xuño de 2012

Para calquera cuestión, pode dirixirse a: `editores@linguamatica.com`

Petición de Artículos

La revista Linguamática pretende cubrir una laguna en la comunidad de procesamiento del lenguaje natural para las lenguas ibéricas. Con este fin, se publicarán artículos que traten el procesamiento de cualquiera de estas lenguas.

Linguamática es una revista completamente abierta. Los artículos se publicarán de forma electrónica y se pondrán a libre disposición de toda la comunidad científica con licencia *Creative Commons*.

Temas de interés:

- Morfología, sintaxis y semántica computacional
- Traducción automática y herramientas de ayuda a la traducción
- Terminología y lexicografía computacional
- Síntesis y reconocimiento del habla
- Extracción de información
- Respuesta automática a preguntas
- Lingüística de corpus
- Bibliotecas digitales
- Evaluación de sistemas de procesamiento del linguage natural
- Herramientas y recursos públicos o cooperativos
- Servicios lingüísticos en la red
- Ontologías y representación del conocimiento
- Métodos estadísticos aplicados a la lengua
- Herramientas de apoyo para la enseñanza de lenguas

Los artículos tienen que enviarse en PDF mediante el sistema electrónico de la revista. Aunque el número de páginas de los artículos sea flexible, se sugiere que no excedan las 20 páginas. Los artículos tienen que identificarse debidamente. Del mismo modo, los comentarios de los miembros del comité científico serán debidamente firmados.

En relación a la lengua usada para la escritura del artículo, se sugiere el uso del portugués, gallego, castellano, vasco o catalán.

Los artículos tienen que seguir el formato gráfico de la revista. Existen modelos LaTeX, Microsoft Word y OpenOffice.org en la página de Linguamática.

Fechas Importantes

- Envío de artículos hasta: 15 de abril de 2012
- Resultados de la selección: 15 de mayo de 2012
- Versión final: 31 de mayo de 2012
- Publicación de la revista: junio de 2012

Para cualquier cuestión, puede dirigirse a: `editores@linguamatica.com`

Petició d'articles

La revista Linguamática pretén cobrir una llacuna en la comunitat del processament de llenguatge natural per a les llengües ibèriques. Així, es publicaran articles que tractin el processament de qualsevol d'aquestes llengües.

Linguamática és una revista completament oberta. Els articles es publicaran de forma electrònica i es distribuiran lliurement per a tota la comunitat científica amb llicència *Creative Commons*.

Temes d'interès:

- Morfologia, sintaxi i semàntica computacional
- Traducció automàtica i eines d'ajuda a la traducció
- Terminologia i lexicografia computacional
- Síntesi i reconeixement de parla
- Extracció d'informació
- Resposta automàtica a preguntes
- Lingüística de corpus
- Biblioteques digitals
- Evaluació de sistemes de processament del llenguatge natural
- Eines i recursos lingüístics públics o cooperatius
- Serveis lingüístics en xarxa
- Ontologies i representació del coneixement
- Mètodes estadístics aplicats a la llengua
- Eines d'ajut per a l'ensenyament de llengües

Els articles s'han d'enviar en PDF mitjançant el sistema electrònic de la revista. Tot i que el nombre de pàgines dels articles sigui flexible es suggereix que no ultrapassin les 20 pàgines. Els articles s'han d'identificar degudament. Igualement, els comentaris dels membres del comitè científic seràn degudament signats.

En relació a la llengua usada per l'escriptura de l'article, es suggereix l'ús del portuguès, gallec, castellà, basc o català.

Els articles han de seguir el format gràfic de la revista. Es poden trobar models LaTeX, Microsoft Word i OpenOffice.org a la pàgina de Linguamática.

Dades Importants

- Enviament d'articles fins a: 15 d'abril de 2012
- Resultats de la selecció: 15 de maig de 2012
- Versió final: 31 de maig de 2012
- Publicació de la revista: juny de 2012

Per a qualsevol qüestió, pot adreçar-se a: `editores@linguamatica.com`

Artilulu eskaera

Iberiar penintsulako hizkuntzei dagokienean, hizkuntza naturalen prozedura komunitatean dagoen hutsunea betetzea litzateke Linguamática izeneko aldizkariaren helburu nagusiena. Helburu nagusi hau buru, aurretik aipaturiko edozein hizkuntzen prozedura landuko duten artikuluak argitaratuko dira.

Linguamática aldizkaria irekia da oso. Artikuluak elektronikoki argitaratuko dira, eta komunitate zientefikoaren eskura egongo dira honako lizentziarekin; *Creative Commons*.

Gai interesgarriak:

- Morfologia, sintaxia eta semantika konputazionala.
- Itzulpen automatikoa eta itzulpengintzarako lagungarriak diren tresnak.
- Terminologia eta lexikologia konputazionala.
- Mintzamenaren sintesia eta ikuskapena.
- Informazio ateratzea.
- Galderen erantzun automatikoa.
- Corpus-aren linguistika.
- Liburutegi digitalak.
- Hizkuntza naturalaren prozedura sistemaren ebaluaketa.
- Tresna eta baliabide publikoak edo kooperatiboak.
- Zerbitzu linguistikoak sarean.
- Ezagutzaren ontologia eta adierazpideak.
- Hizkuntzean oinarrituriko metodo estatistikoak.
- Hizkuntzen irakaskuntzarako laguntza tresnak.

Arikuluak PDF formatoan eta aldizkariaren sitema elektronikoaren bidez bidali behar dira. Orri kopurua malgua den arren, 20 orri baino gehiago ez idaztea komeni da. Artikuluak behar bezala identifikatu behar dira. Era berean, zientzi batzordeko kideen iruzkinak ere sinaturik egon beharko dira.

Artikulua idazterako garaian, erabilitako hizkuntzari dagokionean, honako kizkuntza hauek erabili daiztezke; portugesa, galiziera, gaztelania, euskara, eta katalana.

Artikuluek, aldizkariaren formato grafikoa jarraitu behar dute. "Linguamática" orrian LaTeX, Microsoft Word eta OpenOffice.org ereduak aurki ditzakegu.

Data garratzitsuak:

- Arikuluak bidali ahal izateko epea: 2012eko apirilak 15.
- Hautapenaren emaitzak: 2012eko maiatzak 15.
- Azken itzulpena: 2012eko maiatzak 31.
- Aldizkariaren argitarapena: 2012eko ekainean.

Edozein zalantza argitzeko, hona hemen helbide hau: `editores@linguamatica.com`.

Volume 4, Número 1- Abril 2012

linguaMATICA

ISSN: 1647-0818