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.