Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 10, Número 1 (2018)

ISSN: 1647-0818

linguaMATICA

Volume 10, Número 1 – 2018

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Editorial

Unha vez máis, Linguamática ve a luz neste seu décimo ano grazas á colaboración de todas e de todos, nun esforzo sostido de mantermos aberto e activo este foro de comunicación estabelecido nas nosas linguas peninsulares e no ámbito das tecnoloxías lingüísticas dos nosos idiomas.

Nos últimos anos, a revista Linguamática foise consolidando na comunidade científica do procesamento da linguaxe como un medio prestixioso de difusión dos resultados da investigación e de presentación de proxectos, sendo incluída e reseñada en índices de grande relevancia como Emerging Sources Citation Index, Scopus, ERIH Plus, DBLP, LLBA ou REDIB, entre outros.

Dentro deste esforzo para mantermos e mellorarmos Linguamática como medio de comunicación da nosa comunidade científica e lingüística, para este volume actualizamos o sistema de publicación de Open Journal Systems que sustenta as actividades de edición, publicación e libre acceso aos contidos da revista, ofrecendo así un deseño anovado máis versátil e eficiente que esperamos que sexa do voso agrado.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís,
Universidade de Vigo

Alberto Simões,
Universidade do Minho

Aline Villavicencio,
Universidade Federal do Rio Grande do Sul

Álvaro Iriarte Sanroman,
Universidade do Minho

Ana Frankenberg-Garcia,
University of Surrey

Anselmo Peñas,
Univers. Nac. de Educación a Distancia

Antón Santamarina,
Universidade de Santiago de Compostela

Antoni Oliver González,
Universitat Oberta de Catalunya,

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arantza Díaz de Ilarraza,
Euskal Herriko Unibertsitatea

Arkaitz Zubiaga,
Dublin Institute of Technology

Belinda Maia,
Universidade do Porto

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Université de Caen Basse-Normandie

Gerardo Sierra,
Univers. Nacional Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Irene Castellón Masalles,
Universitat de Barcelona

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Joseba Abaitua,
Universidad de Deusto

Juan-Manuel Torres-Moreno,
Lab. Informatique d'Avignon - UAPV

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Marcos Garcia,
Universidade da Corunha

María Inés Torres,
Euskal Herriko Unibertsitatea

Maria das Graças Volpe Nunes,
Universidade de São Paulo

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Miguel Solla Portela,
Universidade de Vigo

Mikel Forcada,
Universitat d'Alacant

Pablo Gamallo Otero,
Universidade de Santiago de Compostela

Patrícia Cunha França,
Universidade do Minho

Rui Pedro Marques,
Universidade de Lisboa

Susana Afonso Cavadas,
University of Sheffield

Tony Berber Sardinha,
Pontifícia Univ. Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

Artigos de Investigação

Proposta recebida em 8 de Maio 2018 e aceite para publicação em 18 de Julho 2018.

Explorando a Geração Automática de Adivinhas em Português

Exploring the Automatic Generation of Riddles in Portuguese

Hugo Gonçalo Oliveira
CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Portugal
hroliv@dei.uc.pt

Ricardo Rodrigues
CISUC, Instituto Politécnico de Coimbra, Portugal
rmanuel@dei.uc.pt

Resumo

Neste artigo descrevemos um conjunto de experiências realizadas com o objectivo de gerar, de forma automática, adivinhas em português, tendo por base características conhecidas de um conceito. Para além de fazerem sentido, um dos objectivos seria a geração de adivinhas inéditas e, idealmente, com potencial humorístico, nem que por comparação às chamadas "piadas secas". Parte do desafio passou pela identificação de recursos linguísticos adequados ao nosso objectivo, a que se seguiu a definição de um conjunto de regras para os explorar, tendo em conta um conjunto de hipóteses que poderiam potenciar a originalidade e o valor humorístico. Por fim, as adivinhas foram apresentadas sob a forma de pares pergunta-resposta. O artigo foca-se principalmente num tipo de adivinha, cujo modelo de geração é dissecado e resultados são apresentados a título de exemplo. Uma amostra de adivinhas deste tipo foi classificada manualmente por avaliadores humanos que as consideraram geralmente coerentes, moderadamente originais, mas, na sua maioria, com um baixo potencial humorístico. De forma a mostrar que muito pode ainda ser feito para melhorar os resultados anteriores, nomeadamente ao nível do humor, são também apresentadas, de forma breve, outras experiências num estado ainda incipiente.

Palavras chave

geração de adivinhas, humor computacional, geração de linguagem natural, criatividade linguística, criatividade computacional, relações semânticas

Abstract

This article describes several experiments towards the automatic generation of riddles in Portuguese, based on two known features of given concepts. In addition to making sense, one of our goals was to produce novel riddles, ideally, with some humour potential, at least when compared to the so called "piadas secas" (dry jokes). Part of the challenge involved the identification of suitable linguistic resources for this goal, followed by setting rules for their exploitation, given some hypothesis that could increase their novelty and humouristic value. Finally, riddles were rendered as question-answer pairs. The article is mainly focused on a kind of riddles, for which the generation model is dissected, some examples are presented, and a sample was manually classified by human judges, who, overall, rated them as coherent, moderately novel, but still with a low humor potential. In order to show how much can still be done, namely about the humor value, the article also presents, briefly, two other experiments, still in an early stage.

Keywords

riddle generation, computational humor, natural language generation, linguistic creativity, computational creativity, semantic relations

1 Introdução

As adivinhas são uma espécie de jogo ou *puzzle* linguístico, para ser resolvido como forma de entretenimento. Georges & Dundes (1963) definem *riddle*, que aqui traduzimos para adivinha, como "*uma forma tradicional de expressão verbal que contém um ou mais elementos descritivos, e onde alguns podem estar em oposição; aquilo a que os elementos se referem deve ser adivinhado*". As adivinhas estão presentes na maioria das culturas e idiomas, e frequentemente em contextos humorísticos, onde podem ser colocadas como uma pergunta, seguida de uma pausa —de forma a permitir que a audiência pense um pouco— para finalmente revelar a resposta, que pode funcionar como um remate, em inglês, vulgarmente chamado de *punchline*. A resolução de adivinhas requer normalmente criatividade, mas a criativi-

DOI: 10.21814/lm.10.1.268

dade também é necessária na produção de novas adivinhas, diferentes daquelas conhecidas por todos, algo fundamental num acto humorístico original.

Em Portugal, é comum chamar-se "piada seca" a um tipo de piada normalmente curta e nem sempre com muita piada, por ser óbvia ou, pelo menos à primeira vista, não fazer muito sentido. Contudo, este tipo de piadas procura tirar partido do anti-clímax para fazer as pessoas rir. Em 2017, as piadas secas, que podem ser apresentadas sob a forma de adivinha, parecem ter regressado à popularidade, com a sua crescente utilização em programas de televisão, vídeos no YouTube[1], e compilação não só em sítios na web, mas também em livros editados (Pinto et al., 2017). Aliás, actualmente, as pessoas tendem mesmo a generalizar e chamar "piada seca" à maioria das piadas do tipo pergunta-reposta curtas e não muito elaboradas.

O presente trabalho inspira-se fortemente noutros trabalhos científicos que visam a geração automática de adivinhas e de humor verbal, para outras línguas (ver secção 2), mas acaba por ser também motivado pelo já mencionado regresso das piadas secas. O trabalho tem como objectivo a produção automática de adivinhas novas, em português, idealmente com valor humorístico, ainda que, tal como o é para a maioria das piadas, especialmente as secas, esse seja um objectivo difícil de avaliar. A tentativa de gerar humor passará pela introdução de incongruência ou ambiguidade na interpretação das perguntas e/ou respostas das nossas adivinhas, que se baseiam sempre num conceito previamente adquirido. Apesar de tudo, neste fase do trabalho, esta tentativa limita-se a um conjunto de hipóteses que considera conceitos com determinadas caraterísticas, nomeadamente compostos, onde um novo sentido pode ultrapassar a soma do sentido dos seus constituintes, para além de pequenas alterações na ortografia dos anteriores, com impacto no seu som. Tanto quanto sabemos, esta é a primeira vez que um trabalho deste tipo é feito na língua portuguesa.

Apesar de estar nos nossos planos futuros abranger outros tipos de adivinha ou humor verbal, de forma a circunscrever o nosso objectivo, este trabalho começou por explorar alguns recursos linguísticos computacionais para a geração de adivinhas inéditas, tendo por base um pequeno conjunto de modelos de exploração dos recursos e de apresentação das adivinhas. O sistema resultante foi apelidado de Seco.

[1]Veja-se, por exemplo, a série *Batalha das Piadas Secas* em `https://www.youtube.com/user/NaoQueresNada`

Este artigo descreve a exploração dos recursos usados, apresenta alguns resultados, e discute a sua validação manual. Mais especificamente, depois desta introdução, a secção 2 faz uma breve revisão de trabalho relacionado, nomeadamente na geração de adivinhas e de humor verbal. De seguida, na secção 3, descrevemos a abordagem actualmente seguida para produzir adivinhas a partir de um conceito inicial, com duas partes —palavras ou sub-sequências—, e duas características desse conceito. Na secção 4 passamos à enumeração e justificação dos recursos linguísticos utilizados, incluindo a origem dos conceitos iniciais, hipóteses e intuições relacionadas com a sua utilização, e ainda a base de conhecimento semântico usada. A secção 5 é dedicada aos modelos escolhidos para apresentar as adivinhas, ilustrados com alguns exemplos. Das adivinhas geradas, uma amostra foi classificada por colaboradores humanos, através de um serviço de *crowdsourcing*. Os resultados dessa classificação são discutidos na secção 6. Estes sugerem que as adivinhas geradas são coerentes ao nível sintáctico e semântico, são moderadamente originais, mas não têm grande potencial humorístico. Apesar de o humor ser um aspecto subjectivo, isto também mostra que, especialmente quando pensamos em "piadas secas", de forma a acrescentar valor humorístico, é necessário trabalhar este aspecto de forma mais direccionada, e não nos basearmos apenas na escolha dos conceitos, como aconteceu. Para além de analisar as pontuações globais, as adivinhas mais bem pontuadas são reveladas, e os aspectos pontuados são ainda mostrados de acordo com a origem dos conceitos, apresentação, tipo de características usadas, e ainda país origem dos avaliadores. Antes de concluir, na secção 7 apresentamos alguns resultados gerados através de métodos semelhantes aos anteriores, onde procuramos aumentar o potencial humorístico. Estes resultados são ainda muito incipientes e com eles pretendemos mostrar essencialmente que, nesta linha, ainda muito pode ser feito.

2 Trabalho Relacionado

O foco nas adivinhas como tema de investigação não é propriamente recente. Veja-se, por exemplo, o trabalho de Georges & Dundes (1963), que procura precisamente definir o conceito de adivinha (*riddle*). Sobre os fenómenos importantes na criação de uma adivinha, Palma & Weiner (1992) referem que a ambiguidade lexical, resultante, por exemplo, de polissemia ou homofonia, é tão ou mais importante para a criação de adivi-

nhas do que a associação de palavras a determinadas categorias, de acordo com os seus múltiplos sentidos.

Mesmo a geração de adivinhas por programas de computador já é estudada desde a década de 1990, com o trabalho seminal de Binsted & Ritchie (1994), de onde resultou o JAPE, um sistema que gera trocadilhos sob a forma de adivinhas, com base em: um léxico com informação sintática e semântica acerca das palavras e seus possíveis sentidos; um conjunto de esquemas para combinar duas palavras baseadas na sua relação lexical ou fonética; e um conjunto de modelos para apresentar as adivinhas. Para a geração de trocadilhos, a versão inicial do JAPE seguia uma de três estratégias: (i) substituição de sílabas por outras com sons próximos; (ii) substituição de palavras por outras com uma sonoridade próxima; (iii) meta-tese, onde a inversão de sons e palavras sugere uma similaridade de sentido em frases que, de outra forma, seriam entendidas com um significado diferente. Um das adivinhas dadas como exemplo é "*What do you call a murderer that has fibre? A cereal killer*", onde o sistema pode tirar partido das semelhanças entre as palavras *serial* e *cereal*. O STANDUP (Manurung et al., 2008) é um outro sistema que adopta uma abordagem semelhante ao JAPE, mas tem mais cuidado na escolha de palavras, restringidas a vários níveis para uma melhor apresentação e utilização por determinadas audiências, tais como crianças. Para além do inglês, a geração de adivinhas com piada foi também tentada para o japonês (Sjöbergh & Araki, 2007).

Trabalhos mais recentes na geração de adivinhas incluem o TheRiddlerBot (Guerrero et al., 2015), que gera adivinhas acerca de personagens famosas. Depois de seleccionar o nome de uma personagem conhecida de uma base de conhecimento, a geração passa pelas seguintes fases: (i) recuperação de características associadas à personagem; (ii) identificação de personagens análogas, por terem características em comum; (iii) selecção de um modelo de apresentação, com base em algumas das características ou na analogia; (iv) publicação da adivinha numa conta da rede social Twitter; (v) recuperação de nomes alternativos, a partir da Wikipédia. Depois de publicada a adivinha, os utilizadores do Twitter podem tentar responder com o nome da personagem ou um dos seus nomes alternativos. Um exemplo de adivinha para o *Joker*, personagem do filme *Batman*, seria: "*Tell me the name of a person that is the Morpheus of The Dark Knight Rises, is criminal, playful yet cruel, has been seen wearing a purple topcoat.*"

Galvan et al. (2016) também trabalharam na geração de adivinhas, com base na associação de palavras. Dado um conceito, a abordagem proposta passa pelas seguintes fases: (i) recuperação das suas possíveis categorias num tesauro criativo; (ii) recuperação de modificadores associados, e selecção aleatória de um deles; (iii) recuperação de novas categorias, a que o modificador seleccionado também esteja associado; (iv) composição de uma categoria final através da combinação do modificador com uma das novas categorias; (v) utilização de um conceito da categoria final para preencher um modelo textual. Um exemplo de adivinha com resposta "sol" (*sun*) seria: "*What is as hot as soup?*"

Ao contrário do JAPE e do STANDUP, nos dois trabalhos anteriores não há uma preocupação especial com o aspecto do humor. Já um outro trabalho que procura gerar humor verbal (Labutov & Lipson, 2012), incluindo adivinhas, explora a estrutura da rede semântica do ConceptNet. Neste caso, a produção de textos humorísticos tem por base a exploração de caminhos de relações e, para maximizar a incongruência, procura-se alinhar caminhos entre dois conceitos, diferentes mas com alguma sobreposição. Mais precisamente, para adivinhas apresentadas como pergunta-resposta, a pergunta menciona dois conceitos de caminhos diferentes, mas do mesmo domínio (identificados através de *clustering*), enquanto que o conceito na resposta está em um dos caminhos mas pertence a um domínio diferente. Um exemplo de uma adivinha gerada será "*Why is the computer in hospital? Because the computer has virus.*"

Para além de adivinhas, existem trabalhos com vista à geração de outros tipos de humor verbal, incluindo acrónimos com piada (Stock & Strapparava, 2006) —e.g., FBI: *Fantastic Bureau of Intimidation*—, ou mensagens curtas (Valitutti et al., 2013) —e.g., *I've sent you my fart.. I mean 'part' not 'fart'....* Ambos exploram a substituição de palavras para potenciar o humor. As palavras seleccionadas para substituir outras devem começar pela mesma letra ou ter um som próximo e respeitar outras características, como por exemplo serem linguagem tabu (e.g., calão).

Tanto na modalidade verbal como oral, o humor baseia-se geralmente em quatro fenómenos linguísticos (Tagnin, 2005): (i) homonímia — palavras com a mesma escrita e som (e.g., 'banco', de jardim ou instituição financeira); (ii) homofonia — palavras com o mesmo som, mas diferentes grafias (e.g., 'cesta' e 'sexta'); (iii) polissemia — palavras com a mesma grafia e som, mas múltiplos sentidos relacionados (e.g., 'banco', ins-

tituição financeira ou de dados); (iv) paronímia — palavras com grafia e sons próximos (e.g., 'tráfego' e 'tráfico'). Assim, para uma tentativa de humor funcionar, é importante que a audiência seja nativa ou fluente na língua em que uma piada é colocada.

De acordo com Attardo (2008), os textos humorísticos podem classificar-se em três tipos: (i) enredo cómico com uma *punchline*, que é o caso das piadas típicas; (ii) enredo cómico com uma ruptura na narrativa que introduz referências humorísticas acerca do autor ou de que a audiência tenha conhecimento; (iii) enredo cómico com uma complicação central humorística, isto é, uma situação banal onde alguns elementos provocam o riso devido às suas consequências.

Na maior parte dos trabalhos anteriores, as adivinhas são geradas com base em alguma teoria ou através da exploração de bases de conhecimento, que são recursos computacionais essenciais para este fim. Apesar de várias semelhanças com o nosso trabalho, salvo uma excepção para o japonês, todos eles produzem texto em inglês. No nosso caso, o objectivo é gerar adivinhas em português, língua para a qual, tanto quanto sabemos, não existe nenhum trabalho deste tipo. Há, contudo, trabalho anterior, também da nossa autoria, na geração automática de um tipo de humor característico da Internet (Gonçalo Oliveira et al., 2016), baseado numa imagem macro e uma linha de texto, em português, mas que não envolve adivinhas.

3 Abordagem

O principal objectivo do Seco é explorar recursos de conhecimento e linguísticos, em português, para gerar novas adivinhas, com a finalidade de entreter uma audiência, e que poderão, em alguns casos, ter valor do ponto de vista humorístico, devido a um trocadilho inerente. Especificamente, o trabalho descrito neste artigo foca-se em adivinhas baseadas em um conceito e duas características dele, apresentadas como pares pergunta-resposta. Se considerarmos os artefactos produzidos pelo sistema como piadas, de acordo com a classificação de Attardo (2008), o Seco produz humor do primeiro tipo, ou seja, apresenta um enredo (pergunta) e depois remata (resposta). Contudo, estamos ainda a dar os primeiros passos em direcção ao nosso objectivo final e, conforme iremos mostrar mais à frente, o potencial humorístico das adivinhas geradas actualmente ainda é baixo.

De certa forma, a abordagem actual do Seco é fortemente inspirada em trabalhos anteriores que procuram gerar adivinhas, A principal diferença será mesmo a geração de texto em português, o que acaba por implicar a utilização de outros recursos linguísticos (ver secção 4). A principal inspiração será o sistema JAPE (Binsted & Ritchie, 1994), apresentado na secção 2, nomeadamente algumas da adivinhas analisadas pelos seus autores e geradas por este sistema, tais como:

- *What do you get if you cross a zebra with a kangaroo? A striped jumper.*
- *What do you get if you cross a car with a vile substance? Crude oil.*
- *What do you get when you cross a chicken and a power pack? A battery hen.*

Cada uma das anteriores pergunta o resultado do cruzamento entre dois conceitos iniciais e a resposta é um outro conceito, lexicalizado por duas palavras. A principal nuance é que cada uma dessas palavras (características) tem um significado relacionado com um dos conceitos iniciais. Em português também é possível encontrar adivinhas que, de certa forma, se encaixam no modelo anterior, tais como:

- *Qual o resultado do cruzamento de uma galinha com uma cobra? Um pinto longo.*
- *Que resulta do cruzamento entre uma cobra e um ouriço? Arame farpado.*
- *Que resulta do cruzamento entre uma girafa e um papagaio? Um alto-falante.*

De forma a gerar adivinhas deste tipo, o procedimento actual segue uma abordagem de baixo para cima, também ele com algumas semelhanças ao procedimento de geração adoptado pelo JAPE. Mais propriamente, partindo de um conceito inicial, passa-se pelas seguintes fases, até se apresentar a adivinha sob a forma de pergunta-resposta:

1. Extracção de características;
2. Emparelhamento de características;
3. Pontuação e filtragem de características;
4. Apresentação como pergunta-resposta.

No nosso caso, o conceito inicial pode ser uma expressão com duas palavras ou uma palavra única que se pode dividir artificialmente em duas palavras válidas, isto é, que terá duas partes (cw_1 e cw_2), uma para cada palavra. Na primeira fase, cada parte do conceito é considerada individualmente, e as características são recuperadas

a partir de uma base de conhecimento, algumas envolvendo a primeira parte (T_1) e outras a segunda (T_2). As características são representadas através de palavras (fw_1 e fw_2) e obtidas a partir de triplos do tipo *a relacionadoCom b*, onde *a* e *b* são palavras e *relacionadoCom* é o nome de uma relação semântica entre significados de *a* e *b* (e.g., *animal* hiperónimoDe *cão*). Assim, interessam-nos os triplos $t_1 \in T_1$ e $t_2 \in T_2$ que envolvem, respectivamente, cw_1 e cw_2:

- $T_1 : \forall(x, relacionadoCom, y) \in T_1$
  $\rightarrow (x = cw_1 \wedge y = fw_1) \vee (x = fw_1 \wedge y = cw_1)$
- $T_2 : \forall(x, relacionadoCom, y) \in T_2$
  $\rightarrow (x = cw_2 \wedge y = fw_2) \vee (x = fw_2 \wedge y = cw_2)$

Na segunda fase, as características $fw_1 \in T_1$ são emparelhadas com as características $fw_2 \in T_2$. Juntamente com o conceito inicial, cada par $\{fw_1, fw_2\}$ é suficiente para produzir uma adivinha. A figura 1 serve para ilustrar as duas primeiras fases. O conceito tem duas partes, cada uma relacionada com uma característica, sob a forma de uma palavra (fw_1 e fw_2), recuperada da base de conhecimento.

Na terceira fase, as adivinhas são pontuadas automaticamente, onde se procura considerar a comunalidade e representatividade das suas características. Finalmente, na quarta fase, as adivinhas são apresentadas num formato textual, mais propriamente como uma pergunta e uma resposta, de acordo com um modelo pré-definido, onde cada característica é inserida.

4 Recursos Linguísticos

Um dos desafios deste trabalho foi a identificação de recursos computacionais linguísticos que nos pudessem ajudar na geração automática de adivinhas em português, idealmente com algum valor humorístico, tais como as apresentadas na secção 3. No desenvolvimento da versão actual do Seco, foram identificados e explorados recursos para obter os conceitos iniciais (listas), recuperar características (base de conhecimento semântica), pontuar as adivinhas (corpo), e tratar das flexões em diferentes géneros e números para a apresentação da adivinha (léxico morfológico). As secções seguintes descrevem os recursos usados, juntamente com exemplos do seu conteúdo, e motivam a sua seleção.

4.1 Conceitos Iniciais

A geração de adivinhas tem por base um conceito inicial, e a criação de um conjunto de adivinhas deve ser apoiado por uma lista de conceitos que sigam as especificações desejadas. Tendo em conta que o procedimento seguido tira partido de conceitos que se podem dividir em duas partes, expressões com duas palavras pareceram-nos adequadas para este fim. Sendo assim, começamos por explorar os seguintes recursos:

- **Compostos:** uma lista de 180 expressões compostas em português (Ramisch et al., 2016), com instâncias tais como *água doce*, *mau-humor*, ou *primeira mão*.
- **N-Adj:** todos os 289 pares substantivo-adjectivo que ocorrem pelo menos 750 vezes no corpo jornalístico CETEMPúblico (Rocha & Santos, 2000), com instâncias tais como *comunicação social*, *ensino superior* ou *prisão preventiva*.

Os conceitos de ambas as listas encontram-se lexicalizados em duas palavras: um substantivo modificado por um adjectivo. A nossa intuição é que o significado dos compostos disponíveis como tal (Compostos) será mais do que a simples soma dos significados de ambas as palavras, enquanto que a maioria dos pares nome-adjectivo extraídos do corpo terão um sentido mais literal. Mais propriamente, a nossa hipótese é que, se as características estiverem associadas aos significados mais literais, os conceitos da primeira lista irão, potencialmente, resultar em associações mais surpreendentes, podendo mesmo sugerir alguma incongruência e, por isso, serem mais susceptíveis de produzir humor. É certo que, a este respeito, poderíamos ter ido mais longe e, por exemplo, aplicar uma medida para a não-composicionalidade de expressões multipalavra (ver, e.g., Biemann & Giesbrecht (2011)). No entanto, optamos por não o fazer porque, nesta fase, o foco principal do trabalho foi mesmo averiguar o que conseguiríamos obter com uma abordagem simplista, e de que forma os resultados obtidos com os dois tipos de lista se equiparavam.

Tendo em conta que o humor pode resultar de duas palavras com sons próximos (homofonia, paronímia), criámos mais duas listas a partir das anteriores, Compostos-d1 e N-Adj-d1, com expressões em que ambas as palavras são válidas, mas em que uma tem um distância de edição igual 1 (remoção, adição ou substituição), respectivamente para uma expressão nas listas anteriores. Algumas das instâncias geradas incluem:

- *amido oculto*, obtido a partir de *amigo oculto*;
- *véu aberto*, obtido a partir de *céu aberto*;

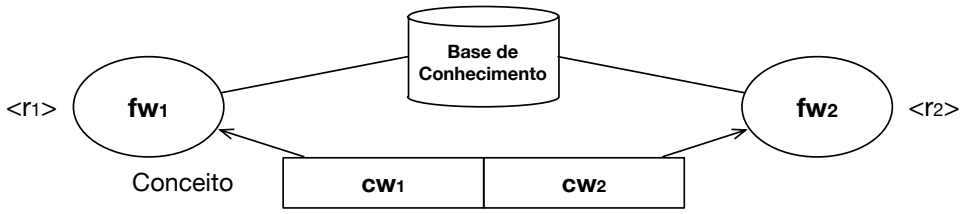

Figura 1: Procedimento para a geração de adivinhas com base em duas caraterísticas de um conceito, obtidas através de uma base de conhecimento.

- *primeiro pano*, obtido a partir de *primeiro plano*.

Aqui, a nossa hipótese foi que, ainda que a interpretação possa ser literal e/ou o composto não seja comum, o paralelismo com um composto original e bem conhecido pode dar origem a algo novo e com algum potencial humorístico.

Uma última lista de conceitos (W1+W2) contém cerca de 900 mil palavras, ou formas, que se encontram léxico LABEL-Lex (Ranchhod et al., 1999) e que podem ser interpretados como amálgamas. Referimo-nos a conceitos do tipo w_1w_2, onde w_1 e w_2 são sequências de caracteres válidas, por também estarem presentes no léxico. Instâncias deste tipo incluem, por exemplo:

- *malabar* = *mala* + *bar*
- *centralidade* = *central* + *idade*
- *restolho* = *resto* + *olho*

A nossa hipótese aqui é que um significado inesperado e, de certa forma, criativo, poderia ser atribuído à palavra, mais uma vez percepcionado como incongruência e ampliando o potencial humorístico.

4.2 Características Consideradas

As características são recuperadas a partir de uma base de conhecimento léxico-semântico obtida a partir de dez recursos lexicais para o português, incluindo redes extraídas de dicionários, wordnets e da ConceptNet (Gonçalo Oliveira, 2018). Mais concretamente, neste trabalho foram usados 45.510 triplos relacionais (*a relacionadoCom b*), correspondentes àqueles que se encontram em pelo menos três dos dez recursos[2]. Cada triplo liga duas palavras, de acordo com os seus significados, mas diferentes sentidos da mesma palavra não são identificados de forma explícita. Ainda que de forma não intencional, isto permite a utilização de significados mais inesperados. A nossa expectativa é que a probabilidade de seleccionar exactamente o mesmo significado que uma palavra tem num termo composto seja baixa, especialmente em compostos onde o significado não resulta da soma dos significados das suas palavras.

Como as relações usadas estão presentes em pelo menos três recursos, elas serão, geralmente, conhecidas e consensuais. Ao levantar essa restrição, seria possível obter mais características, mas não tão imediatas para o público em geral. Haveria também um maior risco de utilizar características incorrectas, porque a maior parte dos dez recursos base foi criada de forma automática ou semi-automática, com um tratamento manual inexistente ou limitado.

Entre as relações cobertas pela base de conhecimento, identificámos um subconjunto de tipos de relação que poderiam ser usados para obter características, dada a sua frequência e adequação à nossa tarefa, considerando também que a maior parte dos conceitos das nossas listas são lexicalizados através de um substantivo e um adjectivo. A tabela 1 mostra os tipos de relação usados para as características, juntamente com a sua frequência na base de conhecimento e com o texto usado para as apresentar nas adivinhas[3].

4.3 Pontuação de Adivinhas

Recorrendo somente ao procedimento anterior, seria possível seleccionar palavras de utilização pouco comum, não muito conhecidas pelo público em geral, ou mesmo características pouco representativas das partes alvo do conceito. Para minimizar estas situações, cada par de características foi pontuado, positivamente e, em alguns casos, negativamente, de acordo com a equação 1, justificada de seguida:

[2] A partir de `http://ontopt.dei.uc.pt/index.php?sec=download_outros`, Triplos relacionais (10 recursos), é possível obter uma lista com todos os triplos nos dez recursos, seguidos do número de recursos em que ocorrem (por exemplo, `fruto HIPERONIMO_DE tomate 3`).

[3] Na verdade, no ficheiro disponível, algumas das relações têm um nome diferente, todo em maiúsculas, também com a categoria gramatical esperada para os argumentos identificados, mas optamos por representar os nomes de relações de uma forma em que essa identificação fosse mais imediata.

#	Relação	arg_1	arg_2	**Apresenação** (r_i)
7,538	adj-sinónimoDe	fw/cw	cw/fw	*o que é fw*
353	adj-antónimoDe	fw/cw	cw/fw	*o que não é/está fw*
4,035	hiperonimoDe	fw	cw	*fw*
590	adj-dizSeSobre-n	cw	fw	*fw*
	adj-dizSeSobre-n	fw	cw	*o que é fw*
100	n-parteDe-adj	cw	fw	*o que é fw*
		fw	cw	*fw*
58	n-parteDe-n	cw	fw	*uma parte de fw*
		fw	cw	*o que tem fw*
49	n-membroDe-n	cw	fw	*um membro de fw*
		fw	cw	*o que tem fw*
46	adj-temQualidade-n	cw	fw	*o que tem fw*
		fw	cw	*o que é fw*
45	n-fazSeCom-n	cw	fw	*a finalidade de fw*
		fw	cw	*o que serve para fw*
110	v-finalidadeDe-n	cw	fw	*a finalidade de fw*
		fw	cw	*o que serve para fw*
1,572	v-causador-n	cw	fw	*fw*
		fw	cw	*o efeito de fw*
101	n-localDe-n	cw	fw	*o que tem fw*
		fw	cw	*o que vem de fw*

Tabela 1: Características extraídas e sua apresentação textual.

- Com o objectivo de favorecer palavras conhecidas pelo público em geral, a pontuação de cada par é proporcional à frequência de cada palavra que representa a característica (fw_1 e fw_2) no corpo CETEMPúblico (α, na equação 2).

- Quando uma das características é uma relação de hiperonímia, o par é penalizado de acordo com o número de hipónimos da característica (β, na equação 3).

A última opção é tomada porque, mesmo com a restrição de usar triplos em pelo menos três recursos, algumas palavras têm muitos hiperónimos (e.g., pessoa, 383; planta, 115; instrumento, 94), o que faz aumentar também o número de respostas possíveis e diminuir a resolubilidade da adivinha. Isto acontece principalmente em redes semânticas que não têm uma taxonomia equilibrada e, por isso, algumas relações de hiperonímia ligam directamente conceitos de uma ordem elevada a conceitos muito específicos (e.g., em vez de ave *hiperonimoDe* animal aquático *hiperonimoDe* animal marinho *hiperonimoDe* pinguim, existe uma ligação directa ave *hiperonimoDe* pinguim).

$$Pontos(par) = \alpha - \beta \quad (1)$$

$$\alpha = \frac{\log(freq(w_1)) + \log(freq(w_2))}{\log(\#maxFreq)} \quad (2)$$

$$\beta = \frac{\log(nHiponimos(w_1))}{\log(\#maxHiponimos)} \quad (3)$$

5 Apresentação de Adivinhas

As adivinhas são apresentadas como pares pergunta-resposta, baseados em um conceito e um par de características extraídas a partir desse conceito, fw_1 e fw_2, por sua vez apresentadas como r_1 e r_2, respectivamente. Aqui, a principal diferença entre fw_1 e fw_2 e r_1 e r_2 é que as últimas podem incluir um artigo antes das características. Esta secção mostra os modelos de apresentação actualmente utilizados, com exemplos para cada um.

A forma de apresentação base para uma adivinha usa o seguinte modelo, motivado pela sua utilização recorrente em adivinhas e trocadilhos:

- *Que resulta do cruzamento entre* $\langle r_1 \rangle$ *e* $\langle r_2 \rangle$ *?* $\langle c \rangle$.

Considere-se, por exemplo, o conceito *direitos humanos*, com o seguinte conjunto de características extraídas:

- *direito* sinónimoDe *liso*
- *direito* sinónimoDe *plano*
- *humano* dizSeSobre *homem*

Ou as seguintes, para o conceito *diapositivo*, dividido em duas partes — *dia* + *positivo*:

- *hora* parteDe *dia*
- *momento* hiperónimoDe *dia*
- *positivo* sinónimoDe *real*
- *positivo* sinónimoDe *confiante*

Os conceitos e características anteriores permitem gerar as seguintes adivinhas:

- *Que resulta do cruzamento entre o que é liso e um homem?*
  direitos humanos.
- *Que resulta do cruzamento entre o que é plano e um homem?*
  direitos humanos.
- *Que resulta do cruzamento entre o que tem uma hora e o que é real?*
  diapositivo.
- *Que resulta do cruzamento entre um momento e o que é real?*
  diapositivo.

O procedimento de geração para a produção da primeira adivinha da lista é ilustrado na figura 2.

Sempre que necessário, recorremos ao LABEL-Lex para identificar o género das características e seleccionar o artigo mais adequado. No seguinte exemplo, é usado o artigo feminino *uma* para a primeira característica e o artigo masculino *um* para a segunda:

- *Que resulta do cruzamento entre uma linguagem e um militar? língua oficial.*

Embora não nos pareça tão crítico, no futuro, o mesmo léxico poderá ser usado para tratar também o número.

Para além da apresentação base, identificámos uma situação específica onde se justifica uma pergunta diferente. Mais propriamente, quando uma das características é uma relação de antonímia, o seguinte modelo é utilizado:

- *Qual é o contrário de* $\langle r_1 \rangle$ $\langle r_2 \rangle$ *?* $\langle c \rangle$*.*

Este modelo funciona para todos os conceitos compostos por um substantivo e um adjectivo, que é o caso da maioria dos conceitos usados. Por exemplo, para *direitos humanos* e *diapositivo*, as seguintes características também foram extraídas:

- *direito* antónimoDe *esquerdo*
- *direito* antónimoDe *torto*
- *humano* antónimoDe *desumano*
- *positivo* antónimoDe *negativo*

Essas características são usadas para gerar as seguintes adivinhas, em que, por ser mais natural, na pergunta, o antónimo é sempre colocado em segundo lugar:

- *Qual é o contrário de homem esquerdo?*
  direitos humanos.
- *Qual é o contrário de homem torto?*
  direitos humanos.
- *Qual é o contrário de plano desumano?*
  direitos humanos.
- *Qual é o contrário de hora negativa?*
  diapositivo.

O procedimento de geração é ilustrado na figura 3 para a última adivinha da lista.

Para além dos modelos anteriores, durante as experiências realizadas, questionámo-nos se as adivinhas funcionariam melhor quando apresentadas tal como descrito — uma pergunta que menciona as duas características e uma resposta que é um conceito potencialmente conhecido — ou se seria preferível inverter a apresentação — uma pergunta simples acerca do significado do conceito inicial, e uma resposta que o explica com recurso às características. De forma a explorar a segunda opção, todas as adivinhas foram também apresentadas através do seguinte modelo:

- *O que significa* $\langle c \rangle$ *?* $\langle r_1 \rangle$ *e* $\langle r_2 \rangle$*.*

Aplicando este modelo nos conceitos *direitos humanos* e *diapositivo*, com as características anteriores, é possível obter as seguintes adivinhas:

- *O que significa direitos humanos?*
  O que é plano e um homem.
- *O que significa direitos humanos?*
  O que é liso e um homem.
- *O que significa direitos humanos?*
  Um homem que não é/está esquerdo.
- *O que significa direitos humanos?*
  Um homem que não é/está torto.
- *O que significa direitos humanos?*
  Um plano que não é/está desumano.

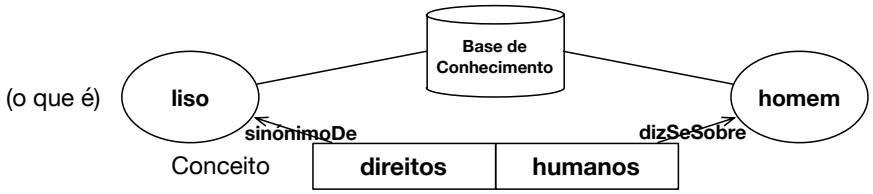

Figura 2: Instanciação do procedimento de geração para o conceito *direitos humanos*, usando o modelo base.

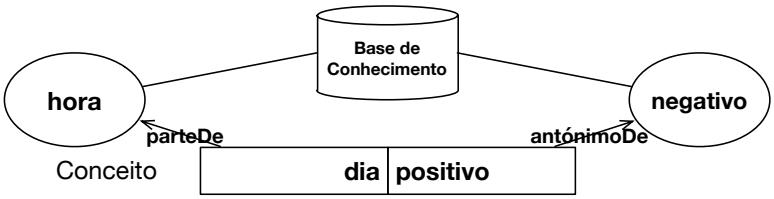

Figura 3: Instanciação do procedimento de geração para o conceito *diapositivo*, usando o modelo para antónimos.

- *O que significa diapositivo?*
  O que tem hora e é/está real.
- *O que significa diapositivo?*
  Um momento que é/está real.
- *O que significa diapositivo?*
  Uma hora que não é/está negativa.

Em busca de algumas conclusões relativamente à apresentação preferível, decidimos gerar as adivinhas seguindo ambos os modelos (Características→Conceito ou Contrário, e Conceito→Características) e, mais à frente, olhar para os resultados de uma avaliação manual independente.

6 Validação

De forma a obter uma opinião próxima daquela do público em geral, uma amostra de adivinhas foi gerada e enviada para a plataforma de crowdsourcing Figure Eight[4] (antigo Crowdflower), onde foi criada uma tarefa (job) a que se deu o nome de *Adivinhas em Português*. Esta tarefa pedia a colaboradores humanos, de Portugal ou Brasil, que classificassem adivinhas de acordo com três aspectos: coerência, novidade e potencial humorístico. Por cada conjunto de respostas completo, cada colaborador teria direito a uma pequena recompensa monetária e, numa tentativa de generalizar as respostas, definiu-se que cada colaborador não poderia classificar mais de 15 adivinhas.

Nesta secção descrevemos os aspectos classificados por cada colaborador, explicamos como a amostra de avaliação utilizada foi gerada, e apresentamos os resultados globais. No entanto, para além da opinião global, esta tarefa foi também desenhada para nos fornecer informação que pudesse ser útil em escolhas a fazer no futuro, com vista a melhorias do sistema. Por isso, também apresentamos e discutimos os resultados organizados de acordo com quatro parâmetros: origem dos conceitos, apresentação das adivinhas, características usadas, e ainda origem dos avaliadores.

6.1 Aspectos classificados

Com o objectivo de manter a tarefa curta, foram colocadas apenas três questões por adivinha, cada uma focada num aspecto diferente, e respondida através de uma escala de Likert de 5 pontos. Com a primeira pergunta, pretendíamos saber se o texto gerado era sintacticamente e semanticamente coerente, e podia ser considerado como uma adivinha. Pedíamos inclusivamente aos colaboradores para, em caso de dúvida, recorrerem a um dicionário. A escala da resposta, explicada aos colaboradores, tinha os seguintes extremos:

- Incoerente (1): a pergunta está mal estruturada, é de difícil interpretação, e a resposta não tem qualquer relação (nem mesmo com um sentido pouco óbvio e depois de pensar um pouco).
- Perfeitamente coerente (5): a pergunta está bem estruturada, tem uma interpretação clara, e a resposta responde efectivamente à pergunta (mesmo que com um sentido menos óbvio e/ou que obrigue a pensar um pouco).

[4] https://www.figure-eight.com/

De seguida, pretendíamos perceber até que ponto a adivinha era nova, surpreendente, e tinha, por isso, valor criativo, o que já envolve algum nível de subjectividade. A pergunta pedia para classificar a originalidade da adivinha numa escala com os seguintes extremos:

- Nada original (1): na sua opinião, a adivinha é demasiado óbvia e nada inovadora.
- Muito original (5): na sua opinião, a adivinha é autêntica / surpreendente e demonstra criatividade.

Por fim, pretendíamos apurar se a adivinha tinha a capacidade de fazer rir e poderia ser usada num contexto humorístico. A pergunta colocada pedia para a classificar precisamente o seu potencial humorístico, numa escala com os seguintes extremos:

- Nenhum potencial (1): na sua opinião, será impossível fazer alguém rir com esta adivinha, por exemplo, por não fazer qualquer sentido ou por ter uma interpretação completamente literal.
- Muita piada (5): na sua opinião, a adivinha tem um grande potencial humorístico e dá vontade de rir.

A classificação por parte dos colaboradores baseou-se apenas no senso comum. Ou seja, para além da descrição das escalas a usar, não foi realizado qualquer tipo de treino nem apresentados exemplos do que seriam adivinhas coerentes, originais e com piada. Ainda que fosse possível escolher um conjunto de adivinhas coerentes e bem formadas, por ser o aspecto menos subjectivo, a novidade e, principalmente, o potencial humorístico têm um elevado nível de subjectividade, para além de que a apresentação de bons exemplos de adivinhas poderia condicionar a classificação da originalidade.

6.2 Geração de amostras

Inicialmente foram geradas adivinhas para cada lista de conceitos referida na secção 4.1, depois ordenadas de acordo com a sua pontuação automática (da maior para a menor). Das anteriores, apenas se mantiveram as 300 adivinhas com maior pontuação para cada lista de conceitos. Garantimos ainda que não havia mais de três adivinhas por conceito inicial. Depois, cada adivinha das anteriores foi apresentada com o modelo Conceito→Características/Contrário e também Características→Conceito, duplicando assim o número de adivinhas usadas. Das anteriores, foi feita uma selecção aleatória de 320 adivinhas, enviada para o Figure Eight, onde cada uma foi classificada por três colaboradores diferentes.

6.3 Resultados globais

A tabela 2 apresenta os resultados globais da validação humana de adivinhas. Para cada aspecto classificado, é mostrado o número de adivinhas para cada resposta possível (entre 1 e 5), a moda, a mediana, e a concordância entre colaboradores. A concordância fica apenas ligeiramente acima dos 50%, o que confirma a subjectividade envolvida. Globalmente, esta validação mostra que a maioria das adivinhas geradas são sintacticamente e semanticamente coerentes — mais de metade foram classificadas com 4 ou 5 —, e pode-se dizer que há uma proporção interessante de adivinhas originais. Por outro lado, os colaboradores não acharam muita piada às adivinhas e consideraram que tinham um baixo potencial humorístico, ao classificarem quase dois terços das adivinhas com 1 ou 2 neste aspecto.

Class.	**Coerência**	**Original.**	**Humor**
1	137	206	444
2	137	231	216
3	216	270	144
4	266	158	86
5	204	95	70
Moda (Mo)	4	3	1
Mediana (Md)	3	3	2
Concord.	51%	59%	58%

Tabela 2: Resultados globais da validação humana.

A correlação de Pearson entre cada aspecto e a pontuação automática de cada adivinha também foi calculada mas estava, para cada aspecto, próxima de zero. Isto pode significar que essa pontuação não reflecte nem a coerência, nem a originalidade, nem o potencial humorístico das adivinhas. Pode também significar que, ao termos criado uma amostra a partir das adivinhas mais bem pontuadas, invalidámos a possibilidade de tirar conclusões acerca desta pontuação. Ou seja, ainda que seja nossa intenção melhorar esta pontuação, terá de ser algo a estudar com mais cuidado, no futuro. A tabela 3 mostra as adivinhas com melhores classificações ao nível da originalidade e do humor.

$\overline{Original}$	$\overline{Humor}$	**Adivinha**	**Pontuação**
4.67	4.67	*Que resulta do cruzamento entre a finalidade de lixa e um treinador? politécnico.*	0.56
4.67	4.00	*Que resulta do cruzamento entre um julgamento e uma sociedade? justiça social.*	0.61
4.33	4.33	*Que resulta do cruzamento entre o que é plano e um homem? direitos humanos.*	0.74
4.33	4.00	*O que significa lei orgânica? uma norma que é um ser.*	0.69
4.33	3.67	*O que significa porto forte? um vinho que é violento.*	0.62
4.33	3.67	*O que significa novo mudo? o que é calado e não é/está antigo.*	0.62
4.33	3.67	*Que resulta do cruzamento entre uma área e o que é marginal? zona ribeirinha.*	0.59
4.33	3.33	*O que significa junção pública? o efeito de juntar que não é/está privado.*	0.68
4.33	3.00	*Que resulta do cruzamento entre o efeito de discutir e o que é manifesto? discussão pública.*	0.62
4.33	2.67	*Que resulta do cruzamento entre o que é chato e uma parte de minuto? segundo plano.*	0.57
4.33	1.00	*O que significa crise económica? o que é crítico e é uma economia.*	0.67
3.00	4.67	*O que significa quadro-negro? uma obra que não é/está branca.*	0.60
4.00	4.00	*Que resulta do cruzamento entre o que é calado e o que é comum? mudo geral.*	0.59
4.00	4.00	*O que significa pronto forte? o que é rápido e é uma força.*	0.71
4.00	4.00	*Que resulta do cruzamento entre o que é mágico e o que é preto? magia negra.*	0.56
3.67	4.00	*O que significa solícito? uma estrela que não é/está ilegal.*	0.54
3.67	4.00	*O que significa politécnico? a finalidade de lixa que é um treinador.*	0.56
3.67	3.67	*Qual é o contrário de pobre velho? novo-rico.*	0.66

Tabela 3: Adivinhas com maior classificação humana na originalidade e humor.

6.4 Resultados parciais

Para cada lista de conceitos iniciais, a tabela 4 mostra o número de adivinhas classificadas juntamente com a moda (Mo) e mediana (Md) dos aspectos avaliados. Verifica-se que a coerência é alta para os termos compostos originais, seguida da lista N-Adj, mas baixa para os novos compostos criados e para as palavras onde a divisão é essencial para forçar um novo significado. A originalidade é comparável para todas a listas, com moda e mediana sempre 3, excepto para a lista N-Adj, com moda 2, o que segue a nossa intuição inicial: o significado dos compostos extraídos do corpo é demasiado literal quando comparado com os termos compostos disponíveis na lista, onde o significado é mais do que uma mera soma dos significados das palavras constituintes.

Para cada modelo de apresentação, a tabela 5 mostra o número de adivinhas classificadas e os respectivos resultados. A coerência é ligeiramente mais alta para o modelo Características→Conceito, mas a principal diferença está na originalidade. Aparentemente, a sensação de novidade é substancialmente menor quando o conceito é referido na pergunta. Isto é uma constatação relevante, a ser considerada no futuro. Contudo, após uma revisão manual dos resultados, não descartamos o impacto negativo de haver demasiados "*que é*" a ocorrer na resposta, por vezes de forma pouco natural. Assim, para já, a nossa decisão será alterar o modelo de apresentação da seguinte forma: quando a primeira característica é um adjectivo e a segunda um substantivo, o modelo passará a considerar que o substantivo é modificado pelo adjectivo. Veja por exemplo esta alteração aplicada a duas das adivinhas anteriores:

- *O que significa direitos humanos?*
  Um homem plano.
- *O que significa direitos humanos?*
  Um homem liso.

Outra alteração que acabamos por realizar, neste caso, no modelo Características→Conceito, foi,

Lista	#	Coerência		Originalidade		Humor	
		Mo	Md	Mo	Md	Mo	Md
Compostos	165	5	4	3	3	1	2
N-Adj	273	4	4	2	3	1	2
Compostos-d1	144	4	3	3	3	1	2
N-Adj-d1	165	4	3	3	3	1	2
W1+W2	213	4	3	3	3	1	1

Tabela 4: Classificação humana de acordo com as listas de conceitos iniciais.

sempre que possível, utilizar os adjectivos como substantivos na pergunta. Aplicado a um dos exemplos anteriores, como a palavra "plano" pode ser utilizada como adjectivo mas também como substantivo, o resultado é:

- *Que resulta do cruzamento entre um plano e um homem?*
  direitos humanos.

Contudo, as alterações anteriores não estão ainda reflectidas nos resultados apresentados da validação.

A tabela 6 mostra, para cada tipo de características extraídas (relações), o número de adivinhas e as suas classificações. Tipos de relação com menos de 10 adivinhas na amostra foram deixados de fora desta tabela. A maioria das adivinhas foi produzida através de sinónimos e hiperónimos, que são também os tipos de relação com mais instâncias na base de conhecimento. Quando agrupados por tipo de relação, os resultados não são conclusivos.

Tal como nos resultados globais, o potencial humorístico é sempre baixo. A originalidade é especialmente baixa para relações de antonímia combinadas com relações do tipo *dizSe-Sobre*. Muitas das adivinhas geradas com essa combinação de características serão demasiado literais, como no seguinte exemplo:

- *Qual é o contrário de vivo público?*
  vida privada.

A combinação entre características do tipo *dizSe-Sobre* com hiperonímia destacou-se com a moda mais alta para a coerência. Segue-se um exemplo de uma das adivinhas obtidas com essa combinação:

- *Que resulta do cruzamento entre uma norma e um ser?*
  lei orgânica.

A originalidade mais elevada foi obtida pela combinação de relações de antonímia e hiperonímia. A seguinte é um exemplo dessa combinação:

- *Qual é o contrário de organismo alto?*
  planta baixa.

Por fim, atendendo às diferenças culturais e linguísticas entre Portugal e Brasil, a tabela 7 apresenta os resultados separados por país. Há algumas diferenças mas, tendo em conta o tamanho da amostra utilizada, o número de avaliadores por adivinha (três) e ainda que os colaboradores de Portugal não terão avaliado exactamente as mesmas adivinhas que os do Brasil, não será possível tirar grandes conclusões. O mais evidente é a menor originalidade atribuída pelos avaliadores de Portugal às adivinhas que têm por base as listas N-Adj (obtida de um jornal português) e Compostos-d1 (que resulta de alterações a compostos recolhidos por uma equipa brasileira), o que tem reflexo no mesmo aspecto a nível global.

7 À procura do humor: exploração de outros modelos de adivinha

Depois das experiências relatadas nas secções anteriores, decidimos explorar outras estratégias para a geração de adivinhas, com algumas semelhanças, mas pequenas diferenças com vista ao aumento do potencial humorístico. Nesta secção descrevemos duas dessas estratégias e revelamos alguns exemplos de adivinhas obtidos em experiências iniciais.

Uma das estratégias passa por criar uma lista semelhante à lista W1+W2, isto é, em que cada entrada tem uma palavra que pode ser dividida em duas partes. Contudo, desta vez, a palavra original sofre pequenas alterações de letras ou sequências que têm um som igual ou com alguma semelhança com a original. Por exemplo, permitindo trocas de sequências como *ce* para *se*, *i* para *e*, *n* para *m*, ou até *on* para *ão* e *ção* para *som*. Apresentam-se alguns exemplos, criados com os mesmos modelos que as adivinhas anteriores, mas depois de gerar palavras com as trocas referidas:

- *Qual é o contrário de orgânico imaginário?*
  sereal.

Modelo	#	Coerência		Originalidade		Humor	
		Mo	Md	Mo	Md	Mo	Md
Características→Conceito	120	4	4	3	3	1	2
Contrário	36	4	3	3	3	1	2
Conceito→Características	166	4	3	1	2	1	1

Tabela 5: Classificação humana de acordo com a apresentação da adivinha.

Características	#	Coerência		Originalidade		Humor	
		Mo	Md	Mo	Md	Mo	Md
sinónimoDe, hiperónimoDe	53	4	3	2	3	1	2
sinónimoDe, sinónimoDe	50	3	3	3	3	1	2
sinónimoDe, dizSeSobre	37	4	4	1, 3	3	1	2
antónimoDe, sinónimoDe	24	4	3	1, 3	3	1	1
antónimoDe, hiperónimoDe	19	4	4	3, 4	3	1	2
causador, sinónimoDe	19	4	4	1, 3	2	1	1
dizSeSobre, hiperónimoDe	17	5	4	3	3	1	2
antónimoDe, dizSeSobre	14	4	3	1	2	1	1.5
partDe, sinónimoDe	12	3, 4	3	3	3	1	1

Tabela 6: Classificação humana de acordo com o tipo de características usado.

Lista	País	#	Coerência		Originalidade		Humor	
			Mo	Md	Mo	Md	Mo	Md
Total	Portugal	390	4	3	2	2	1	2
	Brasil	570	4	4	3	3	1	2
Compostos	Portugal	65	4	4	2	3	1	2
	Brasil	99	5	4	3	3	1	2
N-Adj	Portugal	115	4	4	2	2	1	2
	Brasil	157	5	4	3	3	1	1
Compostos-d1	Portugal	57	2	3	2	2	1	1
	Brasil	87	4	3	3	3	1	2
N-Adj-d1	Portugal	71	3	3	2	3	1	2
	Brasil	94	4	4	3	3	1	2
W1+W2	Portugal	81	1	3	1	2	1	2
	Brasil	132	4	3	3	3	1	1

Tabela 7: Classificação humana de acordo com as listas de conceitos iniciais.

- *Que resulta do cruzamento entre uma sequência e o que é ecológico?*
  genecologia.
- *Qual é o contrário de trabalho doente?*
  deversão.
- *Qual é o contrário de criação imaginária?*
  artereal.
- *Qual é o contrário de história má?*
  bomito.
- *Que resulta do cruzamento entre um imposto e um membro de melodia?*
  obrigasom.
- *Que resulta do cruzamento entre o que é canino e o que é tradicional?*
  cãotradição.
- *Que resulta do cruzamento entre o que é canino e um destino?*
  cãosorte.

Numa outra estratégia, inspiramo-nos numa brincadeira de crianças, onde se procuram gerar falsos antónimos, exemplificada pela adivinha:

- *Qual é o contrário de paixão?*
  mãe-tecto

Para gerar este tipo de adivinhas, recorremos diretamente à lista de conceitos W1+W2, mas com um procedimento focado apenas nos antónimos das duas partes. Seguem-se alguns exemplos de adivinhas deste tipo geradas automaticamente:

- *Qual é o contrário de somali? silêncio-aqui.*
- *Qual é o contrário de bombom? mau-mau.*
- *Qual é o contrário de malcheiroso? bem-fedorento.*
- *Qual é o contrário de causador? efeito-prazer.*
- *Qual é o contrário de reverter? esquecer-carecer.*
- *Qual é o contrário de diapositivo? noite-negativo.*
- *Qual é o contrário de atrocidade? branco-mato.*

De forma a aplicar a ideia original, e sem nenhuma restrição, a norma será gerar palavras inválidas. No futuro, uma forma de melhorar a selecção pode passar por considerar na pontuação a proximidade com palavras do léxico, por exemplo, calculando a distância de edição.

Ambas as estratégias aqui apresentadas requerem, certamente, mais trabalho, tanto ao nível das regras a aplicar e pontuação automática, como também ao nível da validação.

8 Conclusão

Descrevemos neste artigo um conjunto de abordagens baseadas em regras para a geração automática de adivinhas em português, onde se exploram recursos linguísticos disponíveis para esta língua, também eles descritos, juntamente com uma justificação da sua escolha, que teve em vista o aumento do potencial humorístico. Foram ainda apresentados exemplos de adivinhas e os resultados da validação humana de uma amostra, focada na coerência, originalidade e potencial humorístico das adivinhas.

Tanto quando sabemos, este é o primeiro trabalho deste tipo em português. Por se tratar de uma abordagem inicial ao tema, decidimos restringir-nos à geração de um tipo específico de adivinhas — baseadas num conceito conhecido e duas características extraídas, recuperadas de uma base de conhecimento, e apresentadas como um par pergunta-resposta — cuja geração já havia sido realizada por um sistema automático, mas noutras línguas, nomeadamente inglês (Binsted & Ritchie, 1994).

O sistema que envolve as abordagens descritas foi baptizado de Seco. No futuro, pretende-se que produza mais adivinhas deste ou de outros tipos, através da definição de novas estratégias; variações do modelo base de apresentação através de pergunta-resposta; algum tratamento de pequenos aspectos linguísticos, incluindo alguns que, entretanto, já foram tratados com o objectivo de tornar o texto mais natural, ou avaliar a inserção de hífens entre conceitos do tipo w_1w_2, para forçar a interpretação como duas palavras (e.g., *dia-positivo*); e, claro, melhorar o seu potencial humorístico. De forma a mostrar que ainda há muita coisa a experimentar, a penúltima secção deste artigo mostra algumas explorações iniciais, precisamente no sentido de gerar adivinhas com mais piada. Outros tipos a explorar incluem adivinhas do tipo: *Qual é a diferença entre X e Y?*; *O que têm X e Y em comum?*; ou *Qual é o cúmulo de X?*.

Em teoria, seria também possível gerar mais adivinhas do tipo em que nos focámos se fossem exploradas listas alternativas de conceitos iniciais (por exemplo, com verbos), outras bases de conhecimento, ou se fosse considerada a fonética das palavras. O último ponto poderia ser feito com recurso à representação fonética dos grafemas no CETEMPúblico (Veiga et al., 2011). No entanto, não colocando essa possibilidade de parte, acreditamos que em português é possível fazer muita coisa considerando apenas a grafia. No entanto, a partir do momento que tivermos mais adivinhas, torna-se ainda mais crucial conseguir identificar as mais interessantes. Apesar de, actualmente, as adivinhas serem pontuadas de forma automática, os resultados da validação manual sugerem que esta pontuação não está correlacionada com os aspectos avaliados. Algumas ideias a explorar na pontuação automática da originalidade passam por considerar o relacionamento das características usadas — obtido, por exemplo a partir de um modelo distribucional de palavras —, isto é, quanto menor o relacionamento, maior a novidade, sendo que alguma relação terá sempre de de existir com partes do conceito inicial. Pontuar o potencial humorístico será mais desafiante, mas poderão ser exploradas características normalmente utilizadas no reconhecimento automático de humor (e.g., Mihalcea & Strapparava (2006)), tais como ambiguidade ou mesmo a utilização de calão. A utilização desta última em sistemas de geração de humor também não é novidade (Valitutti et al., 2013) e, no caso do português, poderíamos recor-

rer a um dicionário focado precisamente neste registo (Almeida, consultado em 2018). Por exemplo, quando utilizamos a base de conhecimento que inclui todos os triplos, muito maior, o nosso sistema gera a seguinte adivinha, com a presença de calão: "*Que resulta do cruzamento entre um fundo e o que é claro? cu aberto*", obtida com recurso às características *fundo* hiperónimoDe *cu* e *claro* sinónimoDe *aberto*.

Para além dos anteriores, seria importante considerar outros aspectos na pontuação, tais como a resolubilidade, onde se poderia penalizar adivinhas que usam características demasiado genéricas e que, por isso, não restringem suficientemente o número de respostas possíveis. Este aspecto também foi considerado por outros autores (e.g., Labutov & Lipson (2012)) e, no nosso caso, poderia passar por ampliar para outras relações o que já é feito com a penalização de características com um número elevado de hiperónimos.

Por fim, e tal como outros fizeram (e.g., Guerrero et al. (2015), Gonçalo Oliveira et al. (2016)), pretendemos, num futuro próximo, desenvolver um *bot* na rede social Twitter que procurará gerar adivinhas inspiradas na tendências actuais. Como conceitos iniciais, um agente deste tipo poderá usar as tendências, conceitos frequentes em publicações relacionadas com as tendências, ou variações das mesmas.

Referências

Almeida, José João. consultado em 2018. Dicionário aberto de calão e expressões idiomáticas. http://natura.di.uminho.pt/jjbin/dac.

Attardo, Salvatore. 2008. A primer for the linguistics of humor. Em Victor Raskin (ed.), *The Primer of Humor Research*, chap. 3, 101–156. De Gruyter Mouton.

Biemann, Chris & Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. Em *Workshop on Distributional Semantics and Compositionality*, 21–28.

Binsted, Kim & Graeme Ritchie. 1994. An implemented model of punning riddles. Em *12th National Conference on Artificial Intelligence*, vol. 1, 633–638.

Galvan, Paloma, Virginia Francisco, Raquel Hervás & Gonzalo Méndez. 2016. Riddle generation using word associations. Em *10th International Conference on Language Resources and Evaluation (LREC 2016)*, .

Georges, Robert A. & Alan Dundes. 1963. Towards a structural definition of the riddle. *Journal of American Folklore* 76(300). 111––18. doi:10.2307/538610.

Gonçalo Oliveira, Hugo. 2018. A survey on Portuguese lexical knowledge bases: Contents, comparison and combination. *Information* 9(2). 34. doi:10.3390/info9020034.

Gonçalo Oliveira, Hugo, Diogo Costa & Alexandre Pinto. 2016. One does not simply produce funny memes! – explorations on the automatic generation of internet humor. Em *7th International Conference on Computational Creativity*, 238–245.

Guerrero, Ivan, Ben Verhoeven, Francesco Barbieri, Pedro Martins & Rafael Perez y Perez. 2015. TheRiddlerBot: A next step on the ladder towards creative Twitter bots. Em *6th International Conference on Computational Creativity*, 315–322.

Labutov, Igor & Hod Lipson. 2012. Humor as circuits in semantic networks. Em *50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, 150–155.

Manurung, Ruli, Graeme Ritchie, Helen Pain, Annalu Waller, Dave O'Mara & Rolf Black. 2008. The construction of a pun generator for language skills development. *Applied AI* 22(9). 841–869.

Mihalcea, Rada & Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence* 22(2). 126–142.

Palma, Paul de & E. Judith Weiner. 1992. Riddles: Accessibility and knowledge representation. Em *COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics*, 1121–1125.

Pinto, Pedro, João Ramalhinho & Gonçalo Castro. 2017. *O Caderno das Piadas Secas – 500 Tentativas de ter graça*. Manuscrito Editora.

Ramisch, Carlos, Silvio Cordeiro, Leonardo Zilio, Marco Idiart & Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. Em *54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 156–161.

Ranchhod, Elisabete, Cristina Mota & Jorge Baptista. 1999. A computational lexicon of Portuguese for automatic text parsing. Em *SIGLEX99 Workshop: Standardizing Lexical Resources*, 74–80.

Rocha, Paulo Alexandre & Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes (ed.), *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, 131–140.

Sjöbergh, Jonas & Kenji Araki. 2007. Automatically creating word-play jokes in Japanese. Em *Procs. of NL-178*, 91–95.

Stock, Oliviero & Carlo Strapparava. 2006. Laughing with HAHAcronym, a computational humor system. Em *21st National Conference on AI - Volume 2*, 1675–1678.

Tagnin, Stella E. O. 2005. O humor como quebra da convencionalidade. *Revista Brasileira de Lingüística Aplicada* 5(1). 247–257.

Valitutti, Alessandro, Hannu Toivonen, Antoine Doucet & Jukka M. Toivanen. 2013. "Let everything turn well in your wife": Generation of adult humor using lexical constraints. Em *Proceedings 51st Annual Meeting of the Assoc. for Computational Linguistics*, vol. 2, 243–248.

Veiga, Arlindo, Sara Candeias & Fernando Perdigão. 2011. Conversão de grafemas para fonemas em Português Europeu – abordagem híbrida com modelos probabilísticos e regras fonológicas. *Linguamática* 3(2). 39–51.

Proposta recebida em 21 de Março 2018 e aceite para publicação em 16 de Julho 2017.

Estratégias lexicométricas para detetar especificidades textuais

Lexicometric strategies to detect textual specificities

Álvaro Iriarte
Universidade do Minho
Grupo Galabra-UMinho
alvaro@ilch.uminho.pt

Pablo Gamallo
Universidade de Santiago de Compostela
CiTIUSiTIUS-USC
pablo.gamallo@usc.es

Alberto Simões
2Ai Lab – IPCA
Grupo Galabra-UMinho
asimoes@ipca.pt

Resumo

Neste artigo propomo-nos a definir e desenvolver uma estratégia automática para procurar especificidades lexicais dentro de conjuntos de textos utilizando unidades lexicais simples e expressões com várias palavras, ou termos multipalavra (MWE, a sua sigla em inglês).

Propomos uma metodologia para o cálculo da divergência de distribuições de lemas e de MWE que permitirá encontrar, automaticamente, diferenças e semelhanças entre textos não anotados. Esta metodologia poderá ser utilizada para posteriormente identificar grupos de textos sobre os quais se procederá a análises quantitativas e qualitativas semiautomáticas e/ou com intervenção humana.

Num primeiro teste, utilizamos dois textos de especialidade (da área da pediatria) e um texto literário, presumindo que os textos de especialidade deveriam apresentar maiores divergências relativamente ao texto literário do que entre eles próprios. Como os testes feitos mostraram a tendência esperada, decidimos aplicar a mesma metodologia a um segundo grupo de textos (três conjuntos de entrevistas a visitantes da cidade de Santiago de Compostela).

Palavras chave

divergencia de Kullback-Leibler, divergência lexical, lexicometria

Abstract

In this article we propose to to define and develop an automatic strategy to search for lexical specificities within sets of texts using simple lexical units and multiword expressions (MWE).

We propose a methodology for calculating the divergence of lemma and MWE distributions that will automatically find differences and similarities between unlabeled texts. This methodology can be used to subsequently identify groups of texts to which quantitative and qualitative analyzes will be applied (semiautomatically and/or with human intervention).

In a first test, we used two specialized texts (from the area of Paediatrics) and a literary text, assuming that the texts of specialty should present greater divergences with respect to the literary text than among themselves. As the tests that were done showed the expected trend, we decided to apply the same methodology to a second set of texts (three sets of interviews done to visitors in the city of Santiago de Compostela).

Keywords

Kullback–Leibler divergence, lexical divergence, lexicometry

1 Introdução

Dentro das Ciências Humanas e Sociais e mais concretamente nas Humanidades Digitais, há uma necessidade cada vez maior de ter acesso a ferramentas computacionais e estatísticas que permitam detetar semelhanças e diferenças entre grupos de textos (Kilgarriff, 1996) ou medir a riqueza lexical dos mesmos (Tweedie & Baayen, 1998). As análises quantitativas baseadas na distribuição de traços linguísticos são essenciais para o desenvolvimento de trabalhos linguísticos e sociolinguísticos que procuram especificidades e diferenças em textos de natureza e origem diversas. Dentro dos traços linguísticos, têm especial relevância as características lexicais dos textos.

Propomo-nos a definir e desenvolver uma estratégia automática para procurar especificidades linguísticas, nomeadamente especificidades lexicais, dentro de conjuntos de textos. O léxico utilizado por diferentes indivíduos pode diferir substancialmente segundo as propriedades e características dos mesmos, incluindo, como veremos, género, profissão, estudos, etc. O nosso objetivo não é tanto identificar especificidades textuais em relação ao conteúdo específico do texto nem ao seu estilo (tamanho de frases e palavras, etc.), mas sim em relação ao uso de unidades lexicais simples (lemas/palavras) e termos multipalavra (MWE), entendidos aqui como combinações lexicais, n-grams ou cadeias de pala-

DOI: 10.21814/lm.10.1.263

vras (Maia et al., 2008; Stubbs & Barth, 2003) e não no sentido de combinações lexicais restritas, mais frequente dentro da linguistica e da lexicografia (Mel'čuk et al., 1995), combinações lexicais que deveriam funcionar melhor do que as palavras simples ou os lemas, para detetar divergências e convergências textuais, porque deveriam apresentar valores de divergência maiores.

Propomos uma metodologia para o cálculo da divergência de distribuições de lemas e de MWE que permitirá encontrar, automaticamente, diferenças e semelhanças entre textos não anotados. Esta metodologia poderá ser utilizada para posteriormente identificar grupos de textos diferenciados sobre os quais se procederá a análises quantitativas e qualitativas semiautomáticas e/ou com intervenção humana. A identificação destes conjuntos de textos com maior grau de convergência poderá, assim, ser feita sem nenhum tipo de critério ou conhecimento prévio, como o que está a ser utilizado nos testes do presente artigo (*tradução literária vs. texto técnico*; *entrevistas a universitários vs. entrevistas a não universitários*, etc., permitindo assim abordagens, nas análises referidas, com menos riscos de vieses cognitivos ou até de preconceitos.

As nossas hipóteses de partida foram:

1. A divergência de Kullback-Leibler (divergência KL) permite comparar distribuições de palavras e MWE, o que poderá ser usado para comparar automaticamente textos não anotados previamente;
2. O uso de combinações lexicais para detetar divergências e convergências textuais deveria funcionar melhor do que o uso de palavras simples, porque apresentará valores de divergência maiores.

Para levar a cabo o objetivo acima sublinhado, desenvolveremos um método concreto de cálculo da divergência lexical entre textos com base, principalmente, na extração de MWE. Pensamos que esta abordagem poderá acrescentar valor às análises lexicométricas com base na unidade palavra. Uma vez que estas, as palavras, não funcionam como unidades isoladas (Saussure, 1999; Iriarte, 2001), consideramos uma mais-valia para os trabalhos de lexicometria e textometria o facto de ultrapassar a palavra como unidade de análise e descrição linguística. É no mínimo estranho que, com o surgimento de ferramentas informáticas que permitiram abordagens linguísticas mais empiristas, se continue a trabalhar com categorias gramaticais (PoS) já documentadas pelos gregos no ano 100 a.C. (Robins, 1997).

O artigo organiza-se da seguinte maneira: trabalho relacionado (secção 2), descrição do método (secção 3), experiências (secção 4) e conclusões.

2 Trabalho relacionado

O presente artigo presenta uma estratégia para a comparação textual. Existem numerosos métodos cujo objetivo é também a comparação quantitativa entre textos, alguns deles centrados no estilo formal e outros no conteúdo textual. Entre os métodos estilísticos e formais, um dos mais comuns é o que calcula a diversidade lexical a partir de uma família de medidas baseadas na relação entre o número de unidades lexicais e o tamanho total do texto, sendo a mais básica a ratio entre tipos e tokens, chamada *TTR* (McCarthy & Jarvis, 2010; Fergadiotis et al., 2013). Este método permite medir a riqueza lexical dos textos, mas não estabelece a comparação entre eles com base no tipo de unidades lexicais utilizadas. Outros métodos estilísticos centram-se na legibilidade e complexidade dos textos com base no cômputo do tamanho das frases (percentagem de palavras por oração) e das próprias palavras (percentagem de sílabas por palavras) (Loughran & McDonald, 2013), como o que deu lugar ao teste de legibilidade chamado *Flesch-Kincaid*, que mede a dificuldade de um texto para ser compreendido no processo de leitura. Em comparação com os estilísticos e formais, os métodos focados no conteúdo semântico dos textos calculam a similaridade textual com base em modelos distribucionais e *embeddings*, que por sua vez foram inspirados pelos métodos que calculam a similaridade semântica entre orações. As palavras são representadas como vetores de contextos e os documentos (ou pequenos extratos de textos) são modelados como a soma desses vetores. A comparação entre os extratos textuais é levada a cabo mediante medidas de similaridade entre vetores, sendo a mais comum a do *coseno* (Agirre et al., 2016; Mikolov et al., 2013). A diversidade lexical e a legibilidade são estratégias quantitativas muito genéricas centradas no estilo e na forma, enquanto a similaridade distribucional é um método muito mais específico ao focar-se no conteúdo. O método proposto no presente artigo, centrado na divergência lexical mediante lemas e MWE, situa-se a um nível intermédio entre o formalismo estilístico e o conteúdo textual.

Outros trabalhos mais próximos do nosso exploram traços linguísticos concretos —por exemplo uso de pronomes pessoais, de palavras com polaridade, de modificadores nominais, etc.—,

com o intuito de detetar características textuais próprias de grupos sociais: homens/mulheres, jovens, etc. (Argamon et al., 2003)

Por outro lado, e já fora do âmbito do PLN, o trabalho da Rede Galabra[1] centra-se, de maneira especial, em projetos de investigação relacionados com os discursos e práticas culturais na comunidade e os seus impactos, nas suas dimensões económicas, ambientais, socioculturais ou simbólicas (Torres Feijó, 2015; Pazos-Justo et al., 2018)

As nossas responsabilidades, dentro dos referidos projetos, estão relacionadas com o tratamento linguístico (nomeadamente a extração terminológica e a análise lexicométrica) dos *corpora* constituídos: inquéritos, entrevistas, gravações de grupos de discussão e *corpus* documental já catalogado (*vd. infra*).

Utilizando as mesmas orientações metodológicas e o mesmo *corpus*, tentar-se-á replicar alguns resultados dos projetos da Rede Galabra (finalizados e em curso)[2]. Reivindicar a replicação dos resultados na área das Ciências Humanas e Sociais (CHS) é de suma importância, uma vez que é uma prática pouco frequente, o que impede consolidar e validar muitas das nossas pesquisas em CHS como investigação dita científica.

As nossas tarefas consistem em explorar, de maneira automática ou semiautomática, todo o potencial dos inquéritos e entrevistas feitos aos visitantes, bem como a importante base de dados composta pelos produtos literários e culturais já catalogados, mediante o tratamento estatístico e linguístico do *corpus*, neste caso concreto, com a extração terminológica (da base de dados documental já construída e das entrevistas já realizadas), focando, de maneira especial, o que podemos chamar, de maneira genérica, *termos multipalavra (MWE)*, que corresponderão, no nosso trabalho, ao que conhecemos como expressões idiomáticas (*deitar foguetes antes da festa*; *to ask for the moon*), colocações (*ódio mortal*; *bitter hatred*), quase-frasemas (*cartão vermelho*; *black belt*) ou entidades nomeadas (*Santiago de Compostela*, *Cavaleiros da Ordem de Santiago*, *25 de Julho*, etc.) mas também outras combinações lexicais frequentes, não necessariamente restritas (Mel'čuk et al., 1995).

À partida, o uso de MWE deveria ser mais eficaz do que o uso de palavras simples (mesmo que previamente selecionadas) permitindo poder trabalhar com *corpora* não anotados. Por exemplo, em análises posteriores de outros trabalhos do projeto, a palavra *água* será contabilizada na categoria *Gastronomia* quando ocorre em combinações como *beber água*, *água de mesa*, *água mineral*, etc., mas não em *cair na água* ou *água do rio*, por exemplo. Esta última ocorrência, porém, seria contabilizada (erradamente) ao utilizarmos a unidade palavra.

Outro exemplo: formas do adjetivo *caro* que ocorrem nas combinações *preços caros*, *uma cidade cara*, etc. serão contabilizadas na categoria *Economia* ao usarmos MWE com anotação morfossintáctica, evitando a contagem de formas como *cara a cara*, *dar a cara*, *caro colega*, *fazer-se caro*, etc.

3 Descrição do método

Para além da comparação direta de frequências da mesma palavra em dois *corpora*, é possível comparar toda a distribuição das frequências, como um todo. Para isso foi usada a Divergência de Kullback-Leibler (Kullback & Leibler, 1951) que, dada a distribuição de dois corpos diferentes (P_{c_i} e P_{c_j}) pode ser definida por:

$$D_{\mathrm{KL}}(P_{c_i}||P_{c_j}) = \sum_k F_{c_i}(k) \log \frac{F_{c_i}(k)}{F_{c_j}(k)} \quad (1)$$

onde $F_{c_i}(k)$ é a probabilidade (frequência relativa) de palavra k no *corpus* c_i.

A equação 1 permite obter uma medida de quanto a distribuição P_{c_j} se distancia da distribuição P_{c_i}, tomando em conta as probabilidades (ou frequências relativas) das palavras de cada corpus. Para as análises aqui apre-

[1] https://redegalabra.org/

[2] Entre outros:
- Discursos, imágenes y prácticas culturales sobre Santiago de Compostela como meta de los Caminos de Santiago. Projeto de 3 anos de duração financiado pela Subdirección General de Proyectos de Investigación. Dirección General de Investigación Científica y Técnica. Ministerio de Economía y Competitividad. Gobierno de España [Código: FFI2012-35521] (2012-2015); https://redegalabra.org/discursos-imagens-e-praticas-culturais-sobre-santiago-de-compostela-como-meta-dos-caminhos
- Bienestar de la comunidad local através de narrativas y usos culturales: Santiago y el Camino actual [em curso];
- Discursos sobre Santiago de Compostela y el/los Camino(s) de Santiago en la novela española actual (2010) a través de técnicas analíticas digitales: Posibilidades y valor del conocimiento generado [Tese de doutoramento de María Luisa Fernández, orientada por Elias J. Feijó e Roberto Samartim (Grupo Galabra da Univ. de Santiago de Compostela e Grupo Galabra-UMinho)];
- "Narrativas, usos e consumos de visitantes como aliados ou amenaças para o bem-estar da comunidade local: o caso de Santiago de Compostela" (Ref: FFI2017-88196-R), parcialmente subsidiado pelo Ministerio de Industria, Economía y Competitividad do Governo da Espanha no quadro do Programa Estatal de I+D+I Orientada a los Retos de la Sociedad (2018-2021)

sentadas foi usada uma implementação em Perl `Math::KullbackLeibler::Discrete` de um dos autores.[3]

4 Testes

4.1 Objetivos

O nosso objetivo é utilizar a divergência KL para calcular graus de especificidade ou de convergência lexical entre grupos de textos, sem intervenção humana e sem necessidade de trabalhar com *corpora* anotados.

Numa primeira experiência comparamos três textos: dois textos de especialidade (da área da pediatria) e um texto literário (a tradução para o espanhol do romance *Ensaio sobre a Cegueira*, de José Saramago), sabendo à partida que as divergências deverão ser maiores entre qualquer um dos textos de especialidade e o texto literário.

Na segunda experiência, aplicaremos a divergência KL para calcular graus de especificidade ou de convergência lexical entre vários reagrupamentos das entrevistas realizadas a 24 visitantes da cidade de Santiago de Compostela.

Como foi referido, as nossas hipóteses de partida foram:

1. A divergência de Kullback-Leibler (divergência KL) permite comparar distribuições de palavras e MWE, o que poderá ser usado para comparar automaticamente textos não anotados previamente;
2. O uso de combinações lexicais para detetar divergências e convergências textuais deveria funcionar melhor do que o uso de palavras simples, porque deveria apresentar valores de divergência maiores.

Deixamos para futuros trabalhos, com colegas de outras áreas da rede Galabra, as análises relacionais e contrastivas (do ponto de vista qualitativo e quantitativo) destes mesmos lemas e combinações lexicais extraídos das transcrições das entrevistas.

4.2 *Corpus*

A base de dados utilizada pela Rede Galabra disponibiliza o acesso a informação retirada de um *corpus* documental e a um conjunto de inquéritos e entrevistas[4].

Uma vez que, neste momento falta ainda por finalizar o processo de transcrição das entrevistas a portugueses e brasileiros, o trabalho aqui apresentado é realizado apenas sobre 24 entrevistas em castelhano (para além dos três textos utilizados na primeira experiência: dois da área da pediatria e um texto literário).

Com base nos dados disponíveis nos inquéritos, relativos às mesmas pessoas entrevistadas (idade, género, nível de estudos, e autoidentificação como peregrinos ou como turistas), subdividimos as entrevistas nos seguintes subgrupos[5]:

1. Autoidentificação
   - Peregrinos (11 entrevistas)
   - Não peregrinos (13 entrevistas)
2. Nível de estudos
   - Universitários (16 entrevistas)
   - Não universitários (8 entrevistas)
3. Género
   - Mulheres (11 entrevistas)
   - Homens (13 entrevistas)

Antes de calcular o grau de divergência lexical entre os seis conjuntos de entrevistas, foi feito um teste prévio com as frequências dos lemas e das MWE extraídos dos dois textos[6] de especialidade (da área da pediatria). e um texto literário (a tradução para o espanhol do romance *Ensaio*

[3] `https://github.com/ambs/Math-KullbackLeibler-Discrete`

[4] O *corpus* documental disponibiliza aos investigadores do grupo uma base de dados avançada com 560 livros catalogados (Samartim, 2015), procedente de produtos culturais e literários publicados entre 2008 e 2012. O *corpus* foi limitado às produções culturais efetivamente consumidas pelos turistas procedentes da Galiza, Espanha, Portugal e Brasil desde 2008 (Portugal e Brasil são os países de procedência do maior número de visitantes não espanhóis e não comunitários respetivamente).

O *corpus vivo* é constituído por inquéritos e entrevistas a visitantes, comerciantes e comunidade local. Os inquéritos, num total de 2157, foram realizados entre 27/03/2013 e 26/03/2014 a turistas galegos e espanhóis (1323), portugueses (428) e brasileiros (406). Foram gravadas 41 entrevistas a turistas galegos e espanhóis, 59 entrevistas a turistas portugueses e 56 entrevistas a turistas brasileiros.

[5] A divisão por grupos etários será excluída, para o presente trabalho, devido ao reduzido tamanho dos 4 grupos etários estabelecidos no projeto "Discursos, imagens e práticas culturais sobre Santiago de Compostela como meta dos Caminhos": Idade < 30; $30 \leq$ Idade < 45; $45 \leq$ Idade < 69; $70 \leq$ idade.

[6] Cifuentes, Javier & Ventura-Juncá, Patricio (2001). *Manual de Pediatría.* Retrieved January 16, 2018, from `http://botica.com.ve/PDF/6mlped.pdf`;

Cerrolaza, Javier, Mercé, Luis. & Emilio Jardón, Emílio (2008). *1. Consideraciones generales 5 Consideraciones clínicas previas 6.* Retrieved January 16, 2018, from `http://www.espanito.com/1-consideraciones-generales-5-consideraciones-clnicas-previas.html`

sobre a Cegueira, de José Saramago), sabendo à partida, como já referimos, que os dois textos de pediatria deveriam apresentar maior convergência entre si.

Atendendo à Lei de Zifp, as distribuições de palavras baseadas em frequências relativas produzem diferentes escalas de valores com textos de diferentes tamanhos. Para podermos trabalhar com textos de tamanho semelhante e assim comparar os valores usando frequências absolutas, reduzimos os tamanhos dos conjuntos dos textos (as entrevistas e os dois textos utilizados no primeiro teste) ao tamanho do documento mais pequeno de cada grupo. Assim, reduzimos todos os grupos de entrevistas ao tamanho do conjunto de entrevistas feitas a não universitários (64 751 palavras) e o tamanho dos três textos utilizados no primeiro teste, ao tamanho de um dos textos da especialidade de pediatria (25 998 palavras).

O tamanho original dos grupos de entrevistas e dos textos utilizados no primeiro teste são descritos na tabela 1.

texto	*# palavras*
11 entrevistas a peregrinos	71 652
13 entrevistas a não peregrinos	116 285
16 entrevistas a universitários	123 186
8 entrevistas a não universitários	64 751
11 entrevistas a mulheres	92 650
13 entrevistas a homens	102 497
Texto de *Pediatría 1*	35 713
Texto de *Pediatría 2*	25 998
Texto do romance *Ensayo sobre la Ceguera*	107 296

Tabela 1: Tamanhos originais dos textos analisados.

Para os propósitos dos testes aqui realizados, pensamos ser irrelevante o facto de termos cortado os textos de maneira aleatória.

A partir destes conjuntos de textos, foram extraídos os lemas e as MWE com as respetivas frequências e construídas as correspondentes matrizes usadas no cálculo das divergências. A extração de lemas e MWE foi feita com os módulos correspondentes da ferramenta LinguaKit (Gamallo & Garcia, 2017)[7]. O anotador morfossintáctico (que inclui o lematizador) integrado no LinguaKit foi avaliado para três línguas, nomeadamente inglês, português e espanhol, com resultados próximos do estado da arte: ≈ 96% para português e espanhol, e ligeiramente mais baixos (≈ 94%) para inglês (Gamallo et al., 2015; Garcia & Gamallo, 2015). Quanto ao extrator de MWE, foi descrito e avaliado qualitativamente em (Gamallo & Garcia, 2017).

[7] https://github.com/citiususc/Linguakit

Na Tabela 2 apresentamos alguns dados quantitativos relativos aos resultados da extração de lemas e de MWE dos dois textos de especialidade e do texto literário referidos na secção 4.2

	lemas	*lemas (total >1)*	*MWE*	*MWE (total >1)*
Total	4754	2495	6725	798
Ceguera	2209	1371	1226	72
Pediatría 1	2089	1456	2698	419
Pediatría 2	2273	1485	2905	411

Tabela 2: Nº de lemas e MWE extraídos (primeiro teste).

Na Tabela 3 apresentamos dados quantitativos relativos aos resultados da extração de lemas e de MWE da transcrição das 24 entrevistas referidas *supra*.

	lemas	*lemas (total >1)*	*MWE*	*MWE (total >1)*
Total	3907	3370	4271	3145
Mulheres	2203	2062	1526	1280
Homens	2319	2293	1669	1661
Universitários	2211	2165	1679	1553
Não universitários	2204	2124	1658	1441
Peregrinos	2261	2017	1712	1183
Não peregrinos	2281	2281	1687	1687

Tabela 3: Nº de lemas e MWE extraídos das entrevistas.

4.3 Primeiro teste: texto científico *vs.* texto literário

Nesta primeira experiência, a partir das matrizes com as frequências dos lemas e dos MWE extraídos dos dois textos de especialidade e do texto literário referidos na secção 4.2, calculamos o grau de divergência lexical entre os mesmos, presumindo, como dissemos, que os dois textos da área de especialidade deveriam apresentar maior convergência entre si e que as divergências deveriam ser maiores entre estes e o texto literário.

Dado tratar-se de uma divergência, para um par de documentos (d_1, d_2) foi calculada a média das divergências das suas distribuições $D_{\mathrm{KL}}(P_{d_1}||P_{d_2})$ e $D_{\mathrm{KL}}(P_{d_2}||P_{d_2})$.

Os resultados são apresentados em duas colunas, sendo que a segunda corresponde aos resultados de frequências > 1 .

Como veremos, todas as configurações devolvem resultados consistentes entre elas.

4.3.1 Cálculo de divergências usando lemas

Na Tabela 4 apresentamos os resultados da comparação dos dados relativos às listas de frequências dos lemas extraídos de cada um dos textos, comparados dois a dois.

Considerando que a divergência de Kullback-Leibler é nula para duas distribuições idênticas, pode-se concluir, como esperado, que as maiores divergências aparecem entre o texto literário e cada um dos textos de especialidade.

	Lemas	*Lemas(freq > 1)*
Ceguera - Pediatría 1	3,1445	2,8761
Ceguera - Pediatría 2	3,3874	3,0859
Pediatría 1 - Pediatría 2	1,9542	1,6351

Tabela 4: Cálculo de divergências usando lemas (primeiro teste).

4.3.2 Cálculo de divergências usando MWE

Na Tabela 5 apresentamos os resultados da comparação dos dados relativos às listas de frequências das MWE extraídas de cada um dos textos, comparados dois a dois.

Também aqui, como esperado, as maiores divergências aparecem, novamente, entre o texto literário e cada um dos textos de especialidade.

	MWE	*MWE(freq > 1)*
Ceguera - Pediatría 1	13,3517	15,6336
Ceguera - Pediatría 2	13,3193	15,6978
Pediatría 1 - Pediatría 2	12,1861	12,3519

Tabela 5: Cálculo de divergências usando MWE (primeiro teste).

4.4 Segundo teste: entrevistas

Na segunda experiência, a partir das matrizes com as frequências dos lemas e das MWE extraídos da transcrição de 24 entrevistas realizadas, entre 27/03/2013 e 26/03/2014, a 24 pessoas que visitaram a cidade de Santiago de Compostela, calculamos o grau de divergência lexical entre os seis conjuntos de entrevistas já referidos (peregrinos *vs.* não peregrinos; universitários *vs.* não universitários; mulheres *vs.* homens).

Como no caso anterior, dado tratar-se de uma divergência, para um par de documentos (d_1, d_2) foi calculada a média das divergências das suas distribuições $D_{\mathrm{KL}}(P_{d_1}||P_{d_2})$ e $D_{\mathrm{KL}}(P_{d_2}||P_{d_2})$.

Com os conjuntos de entrevistas estudados, o esperado é que as maiores divergências apareçam entre os grupos que se opõem diretamente: entrevistas a mulheres *vs.* entrevistas a homens; entrevistas a peregrinos *vs.* entrevistas a não peregrinos; entrevistas a universitários *vs.* entrevistas a não universitários. Como veremos, todas as configurações devolvem resultados consistentes entre elas.

4.4.1 Cálculo de divergências usando lemas

Na Tabela 6 apresentamos os resultados da comparação dos dados relativos às listas de frequências dos lemas extraídos de cada um dos seis grupos de entrevistas comparados dois a dois.

Pode-se concluir que, ao compararmos os grupos de entrevistas dois a dois, as maiores divergências aparecem entre os grupos que se opõem diretamente: homem–mulher; peregrino–não peregrino; universitário–não universitário.

	Lemas	*Lemas(freq > 1)*
mulheres – homens	**0,5059**	**0,48865**
mulheres – Univer.	0,389	0,3696
mulheres – NãoUniver.	0,2818	0,25865
mulheres – Peregr.	0,3057	0,26615
mulheres – NãoPeregr.	0,41205	0,39725
homens – Univer.	0,16135	0,1539
homens – NãoUniv	0,37225	0,36145
homens – Peregr.	0,4111	0,3841
homens – NãoPeregr.	0,10655	0,10375
Univers. – NãoUniver.	**0,50035**	**0,48785**
Univer. – Peregr.	0,3924	0,36345
Univer. – NãoPeregr	0,13005	0,12535
NãoUniver. – Peregr.	0,30215	0,26945
NãoUniver. – NãoPeregr.	0,36475	0,3567
Peregr. - NãoPeregr.	**0,4699**	**0,44575**

Tabela 6: Cálculo divergências usando lemas.

4.4.2 Cálculo de divergências usando MWE

Na Tabela 7 apresentamos os resultados da comparação dos dados relativos às listas de frequências das MWE extraídas de cada um dos seis grupos de entrevistas, comparados dois a dois.

Neste caso, também se pode concluir que, ao compararmos os grupos de entrevistas dois a dois, as maiores divergências aparecem entre os grupos que se opõem diretamente: homem–mulher; peregrino–não peregrino; universitário–não universitário.

5 Conclusões

As duas experiências apresentadas (a comparação de dois textos de especialidade e de um texto literário —primeiro teste— e a comparação dos três conjuntos de entrevistas a visitantes da cidade de Santiago de Compostela) permitem confirmar que a divergência de Kullback-Leibler (divergência KL) é uma medida robusta porque extrai, em ambos os casos, os valores esperados.

	MWE	*MWE(freq > 1)*
mulheres – homens	**11,7211**	**11,6629**
mulheres – Univer.	9,4818	9,1155
mulheres – NãoUniver.	6,77665	5,88005
mulheres – Peregr.	7,82915	6,5231
mulheres – NãoPeregr.	9,6372	9,38765
homens – Univer.	3,2259	2,86875
homens – NãoUniver.	8,57	8,3035
homens – Peregr.	9,70665	9,1854
homens – NãoPeregr.	2,12425	2,10125
Univers – NãoUniver.	**11,6121**	**11,53955**
Univer. – Peregr	9,62975	8,99675
Univer. – NãoPeregr.	3,1857	2,8462
NãoUniver. – Peregr.	7,84815	6,634
NãoUniver. – NãoPeregr.	8,5075	8,2438
Peregr. – NãoPeregr.	**11,54975**	**11,3894**

Tabela 7: Cálculo de divergências usando MWE.

A configuração que apresenta divergências maiores é a que só toma em conta frequências > 1 e usa lemas.

Portanto, das duas hipóteses de partida:

1. A divergencia de Kullback-Leibler (divergência KL) permite comparar automaticamente textos não anotados;
2. O uso de MWE será mais adequado do que o uso dos lemas porque apresentará valores de divergência maiores;

só se confirmou a primeira, embora se possa afirmar que o uso das MWE também é válido para comparar textos com a divergência KL pois se conseguem resultados igualmente robustos. É preciso também sublinhar que o número de MWE utilizados para calcular as divergências é menor que o de lemas, o que nos leva a inferir que uma extração automática com mais cobertura e exaustividade deveria melhorar os resultados.

No primeiro teste, utilizamos dois textos de especialidade (da área da pediatria) e um texto literário, presumindo que os textos de especialidade deveriam apresentar maiores divergências relativamente ao texto literário do que entre eles próprios. Como as experiências feitas mostraram a tendência esperada, decidimos aplicar a metodologia a um segundo grupo de textos (três conjuntos de entrevistas a visitantes da cidade de Santiago de Compostela).

No segundo teste, baseado em entrevistas, os resultados não só ajudam a confirmar a eficácia da medida de divergência, mas também permitem confirmar a pertinência das categorias socioculturais utilizadas para desenhar as entrevistas, bem como a sua pertinência nas análises qualitativas futuras que pretendemos desenvolver no projeto. Neste sentido, conjeturamos que uma categoria social ou cultural tem traços distintivos diferenciadores se o seu discurso é divergente do de indivíduos doutras categorias.

Deixámos para futuros trabalhos:

1. Procurar estratégias que permitam combinar o uso de formas, lemas e MWE no cálculo de divergências/convergências em textos não anotados.
2. As análises relacionais e contrastivas (do ponto de vista qualitativo e quantitativo) dos lemas e MWE mais relevantes extraídos das transcrições das entrevistas referidas na nota 4.

Agradecimentos

Este trabalho é apoiado pelo projeto *Narrativas, usos e consumos de visitantes como aliados ou amenaças para o bem-estar da comunidade local: o caso de Santiago de Compostela.* Ref: FFI2017-88196-R, parcialmente subsidiado pelo *Ministerio de Industria, Economía y Competitividad* espanhol no quadro do *Programa Estatal de I+D+i Orientada a los Retos de la Sociedad (2018-2021).*

Referências

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau & Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. Em *International Workshop on Semantic Evaluation (SemEval)*, 497–511.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine & Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & Talk* 23(3). 321–346.

Fergadiotis, Gerasimos, Heather H. Wright & Thomas M. West. 2013. Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology* 22(2). 397–408. doi:10.1044/1058-0360.

Gamallo, Pablo & Marcos Garcia. 2017. LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática* 9(1). 19–28. doi:10.21814/lm.9.1.243.

Gamallo, Pablo, Juan Carlos Pichel, Marcos Garcia, José Manuel Abuín & Tomás Fernández-

Pena. 2015. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data. *Procesamiento del Lenguaje Natural* 53. 17–24.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. Em *Languages, Applications and Technologies* CCIS, 65–75. Springer.

Iriarte, Álvaro. 2001. *A unidade lexicográfica. palavras, colocações, frasemas, pragmatemas*: Universidade do Minho. Tese de Doutoramento.

Kilgarriff, Adam. 1996. Why chi-square doesn't work, and an improved LOB-Brown comparison. Em *Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, 169–172.

Kullback, S. & R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1). 79–86. doi:10.1214/aoms/1177729694.

Loughran, Tim & Bill McDonald. 2013. Measuring readability in financial disclosures. *Journal of Finance* doi:10.2139/ssrn.1920411.

Maia, Belinda, Rui Sousa Silva, Anabela Barreiro & Cecília Fróis. 2008. N-grams in search of theories. Em Barbara Lewandowska-Tomaszczyk (ed.), *Corpus Linguistics, Computer Tools, and Applications: State-of-the Art*, 71–84. Peter Lang.

McCarthy, PM & J Jarvis. 2010. Mtld, voc-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 41(2). 381–392.

Mel'čuk, Igor, André Clas & Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Advances in Neural Information Processing Systems*, 3111–3119.

Pazos-Justo, Carlos, María Luísa del Río Araujo & Roberto Samartim. 2018. Políticas culturais e comunidade local: contributos para a análise do caso de santiago de compostela como meta dos caminhos de santiago. Em *Atas do III Congresso Internacional sobre Culturas: Interfaces da Lusofonia*, vol. 7, Instituto de Ciências Sociais da Universidade do Minho. No prelo.

Robins, Robert Henry. 1997. *A short history of linguistics* Longman linguistics library. Longman.

Samartim, Roberto. 2015. Bases de dados para o estudo da cultura: apresentação do catalogador e possibilidades de abordagem sobre o corpus documental do projeto caminho de santiago. Em *Estudos da AIL sobre teoria e metodologia*, vol. 2, 115–125. AIL Editora.

Saussure, Ferdinand. 1999. *Curso de linguística geral*. Lisboa: Dom Quixote.

Stubbs, Michael & Isabel Barth. 2003. Using recurrent phrases as text type discriminators: a quantitative method and some findings. *Functions of Language* 10(1). 61–104.

Torres Feijó, Elias. 2015. Identity sustainability, identity affectivity, and the ithaca traveler: Conceptual tools for measuring and modeling tourism as an opportunity. Em Gabriel R. Ricci (ed.), *Travel, Tourism and Identity, Culture & Civilization*, vol. 7, 143–162. Transaction Publishers.

Tweedie, Fiona J. & Harald R. Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities* 32(5). 323–352.

Projetos, Apresentam-se!

Proposta recebida em 7 de Maio 2018 e aceite para publicação em 16 de Julho 2018.

PLN.pt: Processamento de Linguagem Natural para Português como um Serviço

PLN.pt: Natural Language Processing for Portuguese as a Service

Nuno Ramos Carvalho
United Nations University (UNU-EGOV)
ramos.de.carvalho@unu.edu

Alberto Simões
2Ai Lab / IPCA
asimoes@ipca.pt

Resumo

As técnicas da área de Processamento de Linguagem Natural (PLN) são cada vez mais utilizadas para enriquecer aplicações nas mais diversas áreas. As ferramentas que implementam ou apoiam o desenvolvimento destas técnicas podem ser complexas de manter e explorar, sendo por vezes necessário conhecimento específico do domínio.

Este artigo introduz o projeto PLN.pt, uma plataforma online que disponibiliza um conjunto de ferramentas para PLN como um serviço *web* (REST API), orientado principalmente para a língua portuguesa.

Palavras chave

PLN, REST, API, serviço web, português

Abstract

Natural Language Processing (NLP) techniques are often used to enrich applications in several areas. Tools that implement or support the development of these techniques can be complex to maintain and exploit, and specific domain knowledge is usually required.

This paper introduces the PLN.pt, an online platform that enables a set of tools for natural language processing as a web-service (REST API), mainly focused on the portuguese language.

Keywords

NLP, REST, API, web service, Portuguese

1 Introdução

Se até há bem pouco tempo a investigação em Processamento de Linguagem Natural (PLN) era essencialmente académica, e poucas eram as áreas onde tal investigação era aplicada comercialmente ou com objetivos comerciais ou industriais, recentemente esta área tem vindo a ter um interesse crescente na indústria. Se áreas como a tradução automática Rychtyckyj (2006) ou técnicas como a indexação de documentos Frakes & Baeza-Yates (1992), eram já populares, cada vez mais surgem novas áreas de interesse, como a compilação automática de resumos sobre notícias e a sua agregação Mani & Maybury (1999) ou a análise de redes sociais para extrapolar a relevância e aceitação dos produtos comercializados Liu (2015).

Atualmente, não existem soluções "*out of the box*," que se possam adquirir e que implementem, como caixa negra, o que se pretende. Este tipo de aplicação é, ainda, desenvolvida caso a caso, e na sua maioria, em parcerias académicas.

A implementação destas soluções requer conhecimento específico de PLN, quer para desenhar a sua arquitetura, quer para escolher e usar as ferramentas relevantes. Se por um lado adquirir o conhecimento científico necessário é complicado, é igualmente trabalhosa a instalação e configuração das ferramentas disponíveis para as diferentes tarefas necessárias.

É neste campo que o PLN.pt pretende atuar, disponibilizando um conjunto de ferramentas de diferentes níveis de complexidade, que permitam a execução de tarefas, desde as tarefas simples de *atomização*, *segmentação* ou *anotação morfossintática* (PoS) até tarefas mais complicadas como a *análise sintática*, a *construção de árvores de dependências* e, no futuro, até mesmo a *classificação de sentimento* ou a *sumarização*.

Na verdade, qualquer aplicação que tire partido de técnicas de PLN necessita de uma pilha de ferramentas específicas de PLN, que usadas em conjunto permitam criar e explorar os recursos necessários para possibilitar as funcionalidades pretendidas. Apesar de haver uma série de pilhas de ferramentas atualmente disponíveis, como sejam o FreeLing Padró & Stanilovsky (2012), o NLTK Loper & Bird (2002), ou o OpenNLP Baldridge (2005), estes nem sempre são fáceis de instalar e de manter, já que por ve-

DOI: 10.21814/lm.10.1.267

zes surgem conflitos de requisitos, dependências entre bibliotecas, etc., que levam a que seja gasto muito tempo na sua configuração e que, muitas vezes, tem de ser replicada sempre que se pretende implementar uma nova aplicação. Além disso, em alguns casos, é necessário algum conhecimento específico sobre o domínio para tirar o melhor proveito de cada uma das ferramentas.

A abordagem que apresentamos tenta colmatar estas situações e dificuldades, e consiste na criação de uma plataforma central onde toda a informação é processada e os recursos são disponibilizados através de uma série de operações pré-definidas independentes do contexto. Isto permite que outras ferramentas possam usar, de uma forma expedita, estas funcionalidades, bastando para isso a realização de pedidos HTTP e serem capazes de processar resultados na notação JSON.

Este artigo pretende introduzir a plataforma e a sua arquitetura (apresentadas na Secção 2), a API atualmente disponibilizada e exemplos de uso (Secção 3), e as bibliotecas desenvolvidas que permitem a abstração do serviço REST (Secção 4). Finalmente, são apresentadas algumas considerações finais e propostas para trabalho futuro (Secção 5).

2 Plataforma e Arquitetura

A plataforma pode ser vista como um encapsulamento de bibliotecas ou aplicações existentes, e é composta por duas partes:

1. As ferramentas específicas, que implementam as várias funcionalidades disponibilizadas;
2. A transformação dos pedidos REST em parâmetros para as respetivas ferramentas e o tratamento do resultado, transformando-o numa estrutura normalizada em JSON.

2.1 Ferramentas

A plataforma não pretende ser uma implementação de uma ou mais funcionalidades referentes a tarefas de PLN, mas pretende sim disponibilizar o acesso, uniformizado, a um conjunto de ferramentas já existentes.

Para isso, foi necessário escolher um conjunto inicial de funcionalidades a disponibilizar e as ferramentas responsáveis pela sua implementação.

Como foi referido na introdução, o primeiro objetivo do PLN.pt foi o de disponibilizar as ferramentas necessárias para uma pilha de processamento de linguagem natural, que permita o tratamento inicial do texto, com os processos de atomização, segmentação, análise morfológica e análise sintática. Sendo o objetivo a disponibilização para a língua portuguesa, optou-se pela biblioteca FreeLing Padró & Stanilovsky (2012); Simões & Carvalho (2012). Para além de ser de código aberto e gratuito, suporta várias línguas. Também a sua proximidade geográfica e cultural com a língua portuguesa levou a que fosse uma escolha natural.

Embora o FreeLing inclua um analisador morfológico, o carregamento de dados é algo demorado, pelo que, embora seja suficientemente adequado para o processamento de blocos de texto, não é a ferramenta ideal quando se pretende obter as possíveis categorias e propriedades morfológicas de palavras individuais. Nesse sentido, foi também incluído o acesso ao jspell Simões & Almeida (2002), uma ferramenta mantida por um dos autores.

Finalmente, para a disponibilização de um serviço de construção de árvores de dependências, optou-se pelo uso do SyntaxNet Andor et al. (2016). O FreeLing também inclui um módulo de árvores de dependências, mas não disponibiliza um modelo para a língua portuguesa, e num projeto anterior de um dos autores já tinha sido criado um modelo para a SyntaxNet, razão pela qual foi escolhido.

Cada uma destas aplicações é executada, de acordo com uma série de opções bem definidas por omissão, e com os dados necessários, pela componente de interligação do serviço REST.

2.2 Serviço REST

A segunda componente do PLN.pt é então uma aplicação que implementa um serviço REST via HTTP, que implementa a API que permite aceder a todas as ferramentas disponibilizadas de uma forma simples e rápida. Esta aplicação está disponível no GitHub[1], sob uma licença de código aberto, e pode ser executada em qualquer sistema desde que as ferramentas necessárias estejam disponíveis.

De uma forma genérica, esta aplicação executa os seguintes passos para execução de cada pedido:

1. Receber um novo pedido através de um GET ou POST, e que respeite os parâmetros definidos pela API.
2. Utilizar as ferramentas disponíveis para executar a operação e obter um resultado bruto.

[1] https://github.com/nunorc/PLN-PT-api

3. Normalizar o resultado bruto e encapsular o resultado em formato JSON.
4. Devolver o resultado no formato correto para a aplicação cliente.

3 API e Serviços

A API disponibiliza uma série de serviços em que, apesar de serem baseados em ferramentas distintas, os resultados são sempre devolvidos usando uma estrutura coerente.

Esta secção lista os vários serviços, apresentando para cada um o tipo de método HTTP usado (GET ou POST), o endereço do serviço (*endpoint*), o corpo do pedido, sempre que este seja do tipo POST, e o resultado obtido.

3.1 Atomização

Serviço baseado no FreeLing, é responsável por dividir o texto num conjunto de átomos ou *tokens*. O resultado é uma lista de *strings*, em que cada posição corresponde a uma palavra ou a um átomo (pontuação, números, endereços web, etc).

Habitualmente, é útil o uso de segmentação e atomização sobre o mesmo texto, pelo que é possível usar a opção `sentences=1`, obtendo-se uma lista em que cada posição corresponde a uma frase ou segmento, contendo, por sua vez, uma lista dos átomos dessa frase/segmento.

POST	`http://api.pln.pt/tokenizer`
Corpo	`A Maria tem razão.`
Resposta	`["A","Maria","tem","razão","."]`

3.2 Etiquetação Morfossintática

Usa o FreeLing para, além de realizar segmentação e atomização, anotar o texto com *part-of-speech*. O resultado inclui, para cada palavra, o seu lema, etiqueta POS e confiança associada a essa etiquetação.

POST	`http://api.pln.pt/tagger`
Corpo	`A Maria tem razão.`

Resposta

```
[{"pos":"DA0FS0","form":"A","prob":"0.
    675415","lemma":"o"},{"lemma":"maria"
    ,"pos":"NCFS000","form":"Maria","prob
    ":"1"},{"lemma":"ter","form":"tem","
    prob":"0.999287","pos":"VMIP3S0"},{"
    lemma":"razão","pos":"NCFS000","form"
    :"razão","prob":"0.65"},{"lemma":".",
    "pos":"Fp","form":".","prob":"1"}]
```

3.3 Análise de Dependências

Um *parser* de dependências é capaz de analisar a estrutura gramatical de uma frase e gerar uma árvore (ou representação semelhante) das relações binárias entre os elementos léxicos, normalmente chamadas dependências (e.g., sujeito, predicado).

Esta análise é efetuada pelo SyntaxNet, o resultado é normalizado para uma estrutura bem definida, os nomes das relações usados (incluindo os pormenores que cada uma representam) estão disponíveis na coleção do Universal Dependencies[2]. O resultado é uma lista de elementos, em que para cada átomo (*token*) da frase são indicados uma série de propriedades, como por exemplo, a relação de dependência e o átomo pai.

POST	`http://api.pln.pt/dep_parser`
Corpo	`A Maria tem razão.`

Resposta

```
[{"upostag":"DET","deps":"_","head":"2","
    lemma":"_","xpostag":"art||F|S","id":
    "1","feats":"Definite=Def|Gender=Fem|
    Number=Sing|PronType=Art|fPOS=DET++
    art||F|S","form":"A","misc":"_","
    deprel":"det"},{"upostag":"PROPN","
    deps":"_","lemma":"_","head":"3","
    xpostag":"prop|F|S","misc":"_","
    deprel":"nsubj","id":"2","feats":"
    Gender=Fem|Number=Sing|fPOS=PROPN++
    prop|F|S","form":"Maria"},(...)]
```

3.4 Análise Morfológica

A análise morfológica, tal como já foi referido, é realizada pelo jspell. O serviço recebe uma palavra e apresenta uma lista de análises realizadas, incluindo o lema, categoria gramatical, e propriedades de *Part-of-Speech*.

POST	`http://api.pln.pt/word_analysis`
Corpo	`gato`

Resposta

```
[{"rad":"gatinhar","CAT":"v","TR":"i","T"
    :"p","N":"s","P":"1"},
 {"G":"m","N":"s","GR":"dim","CAT":"nc","
    rad":"gato"}]
```

4 Bibliotecas

A API pode ser utilizada a partir de qualquer linguagem de programação, desde que esta seja

[2]Disponível em: http://universaldependencies.org/ (último acesso: 05-05-2018).

capaz de realizar pedidos HTTP. Isto permite que até se possam implementar ferramentas que executem sobre um *browser*, realizando pedidos à API.

De forma a facilitar o desenvolvimento de aplicações em Perl ou em Python, são disponibilizadas bibliotecas para estas linguagens, que encapsulam os pedidos REST, permitindo ao utilizador a manipulação dos resultados sem ter de lidar com o formato JSON.

De seguida faz-se uma pequena demonstração de utilização destas bibliotecas.

4.1 Perl

A biblioteca disponibilizada para Perl, chamada `PLN::PT`[3], implementa um objeto que disponibiliza um método para cada uma das funcionalidades da API, devolvendo os resultados em estruturas de dados nativas da linguagem.

```
use PLN::PT;

my $pln = PLN::PT->new('http://api.pln.pt');
my $text = 'A␣Maria␣tem␣razão.';
my $tokens = $pln->tokenizer($text);

# ['A','Maria','tem',razão','.']
```

Listagem 1: Exemplo de script Perl.

A Listagem 1 ilustra a utilização da biblioteca para dividir uma frase em palavras. Começa por criar um objeto `$pln`, instância da classe `PLN::PT`, para depois invocar o método `tokenizer` passando-lhe, como argumento, o texto. O método devolve uma estrutura (referência para lista de palavras) em Perl.

4.2 Python

A biblioteca disponibilizada para Python, chamada `plnpt`, está disponível no GitHub[4].

O funcionamento desta biblioteca é análogo à sua congénere em Perl: um objeto é implementado que disponibiliza uma série de métodos para aceder a cada uma das operações da API.

```
import plnpt

text = 'A␣Maria␣tem␣razão.';
tokens = plnpt.tokenizer(text)

# ['A', 'Maria', 'tem', 'razão', '.']
```

Listagem 2: Exemplo de script Python.

O Exemplo 2 ilustra a utilização da biblioteca de Python, para dividir uma frase em *tokens* (palavras). É chamada o método `tokenizer` da biblioteca, e passado como argumento o texto a processar, e o resultado é uma lista de *tokens* em Python.

5 Conclusão e Trabalho Futuro

Este artigo descreve o desenvolvimento de um serviço REST que disponibiliza uma série de operações comuns usadas em aplicações da área de PLN, especialmente orientadas para o português, através de uma API. As pilhas de ferramentas e *toolkits* necessários para estas operações podem ser complexos de providenciar e manter, esta aproximação elimina completamente este esforço, e permite o desenvolvimento mais rápido e quase independente das aplicações.

A plataforma está disponível em `http://pln.pt` e, apesar de contar ainda com apenas um número limitado de operações, já se mostrou bastante útil em várias aplicações, em áreas diversas. O seu uso reduz, efetivamente não só o esforço de novas aplicações, como deixa de ser necessário conhecimento especializado para a utilização destas ferramentas.

Como trabalho futuro pretende-se a adição de novas ferramentas e bibliotecas, permitindo assim adicionar novas funcionalidades à API. Será, posteriormente, necessário garantir a sobrevivência da aplicação, aplicando restrições de uso, assim que a quantidade de utilizadores assim o exija.

Agradecimentos

Este artigo foi parcialmente desenvolvido no âmbito do projecto "SmartEGOV: Harnessing EGOV for Smart Governance (Foundations, methods, Tools) / NORTE-01-0145-FEDER-000037", cofinanciado pelo Programa Operacional Regional do Norte (NORTE 2020), através do PORTUGAL 2020 e do Fundo Europeu de Desenvolvimento regional (FEDER).

Os autores agradecem ainda ao Mário Peixoto e aos revisores da Linguamática pela revisão e comentários que ajudaram a melhorar o artigo.

Referências

Andor, Daniel, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov & Michael Collins. 2016. Globally normalized transition-based neural networks. *CoRR* abs/1603.06042. `http://arxiv.org/abs/1603.06042`.

[3] `https://metacpan.org/release/PLN-PT`
[4] `https://github.com/nunorc/plnpt`

Baldridge, Jason. 2005. The OpenNLP project. http://opennlp.apache.org (Último acesso: 17-10-2017).

Frakes, William Bruce & Ricardo Baeza-Yates. 1992. *Information retrieval: Data structures & algorithms*. Prentice Hall Englewood Cliffs, New Jersey.

Liu, Bing. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Loper, Edward & Steven Bird. 2002. NLTK: The natural language toolkit. Em *ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, 63–70.

Mani, Inderjeet & Mark T. Maybury. 1999. *Advances in automatic text summarization*, vol. 293. MIT Press.

Padró, Lluís & Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. Em *Language Resources and Evaluation Conference (LREC 2012)*, .

Rychtyckyj, Nestor. 2006. Machine translation for manufacturing: A case study at ford motor company. Em *18th Conference on Innovative Applications of Artificial Intelligence - Volume 2* IAAI'06, 1728–1735.

Simões, Alberto & Nuno Carvalho. 2012. Desenvolvimento de aplicações em Perl com FreeLing 3. *Linguamática* 4(2). 87–92.

Simões, Alberto Manuel & José João Almeida. 2002. `jspell.pm` — um módulo de análise morfológica para uso em processamento de linguagem natural. Em *Actas da Associação Portuguesa de Linguística (APL2001)*, 485–495.

http://www.linguamatica.com/

linguaMATICA

Artigos de Investigação

Explorando a Geração Automática de Adivinhas em Português
Hugo Gonçalo Oliveira & Ricardo Rodrigues

Estratégias Lexicométricas para Detetar Especificidades Textuais
Álvaro Iriarte, Pablo Gamallo & Alberto Simões

Projetos, Apresentam-se!

PLN.pt: Processamento de Linguagem Natural para Português como um Serviço
Nuno Ramos Carvalho & Alberto Simões