Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 11, Número 1 (2019)

ISSN: 1647-0818

linguaMATICA

Volume 11, Número 1 – 2019

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Editorial

Encetamos o undécimo ano da Linguamática cun volume de cinco artigos cunha maquetación adaptada aos estándares máis recentes de edición científica académica. O novo deseño para as contribucións da revista salientará mediante recursos tipográficos específicos o código ORCid dos asinantes dos artigos e o identificador DOI das referencias bibliográficas incluídas.

Outra mellora menos evidente para os lectores da revista atinxe ao proceso de revisión por pares seguido para a selección dos artigos aceptados para publicación. A partir deste volume, o sistema de revisión será continuo, isto é, as propostas de artigos enviadas á revista serán enviadas aos revisores do comité científico no momento en que se reciban, facilitando así unha xestión máis fluída do proceso de avaliación.

Esperamos que estes cambios contribúan a consolidar a Linguamática como revista científica de referencia e foro privilexiado de comunicación no ámbito das tecnoloxías lingüísticas das nosas linguas peninsulares.

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

Bruno Martins,
Instituto Superior Técnico

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

Mário Rodrigues,
Universidade de Aveiro

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

Patricia Martin Rodilla
Universidade de Santiago de Compostela

Ricardo Rodrigues
CISUC / Instituto Politécnico de Coimbra

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 Abril 2018 e aceite para publicação em Junho 2019.

Uma Utilidade para o Reconhecimento de Topónimos em Documentos Medievais

A Tool for Toponym Recognition in Medieval Documents

Xavier Canosa
CiTIUS / Univ. de Santiago de Compostela
canosarodrigues@gmail.com

Pablo Gamallo
CiTIUS / Univ. de Santiago de Compostela
pablo.gamallo@usc.es

Xavier Varela
ILG / Univ. de Santiago de Compostela
xavier.varela@usc.es

José Ángel Taboada
CiTIUS / Univ. de Santiago de Compostela
joseangel.taboada@usc.es

Paulo Martínez Lema
ILS / Univ. de Santiago de Compostela
paulo.martinez.lema@edu.xunta.es

Marcos Garcia
Grupo LyS, Dpto. de Letras / Univ. da Corunha
marcos.garcia.gonzalez@udc.gal

Resumo

Este artigo apresenta o método de construção duma ferramenta para a anotação de entidades geográficas mencionadas em textos medievais. A nova ferramenta foi desenvolvida a partir dos módulos de língua contemporânea do LinguaKit, pacote multilingue de ferramentas de PLN. Uma coleção de corpora anotados manualmente serviu de recurso para elaborar uma lista de topónimos medievais (*gazetteers*) e observar padrões para a melhora e implementação de novas regras de reconhecimento dos nomes de lugar. Depois da lista de entidades geográficas, os ativadores contextuais (*triggers*) foram o recurso determinante na melhora da abrangência. Para o produto final, fizeram-se também ajustes menores na procura de recolher os elementos mais comuns do léxico e os contextos gramaticais das entidades geográficas mencionadas. Ainda que muito trabalho fica por fazer na elaboração de listas para entidades não geográficas, na construção dum modelo de língua medieval e um lexicon específico, o novo módulo pode ser utilizado para anotar textos e mostra uma melhora significativa a respeito dos módulos previamente existentes.

Palavras chave

entidades geográficas mencionadas, NERC, toponímia

Abstract

This paper describes a method to build a tool aimed at recognizing geographical named entities in medieval texts. The new tool has been developed using the corresponding modules for contemporary languages contained in LinguaKit, a suite of NLP tools. A collection of manually annotated corpora served as a resource to build a gazetteer of medieval toponyms and find patterns to improve and implement new rules for the recognition of place names. In addition to the gazetteer, a list of triggers was the most determinant factor to improve recall. Final adjustments considered the most frequent terms of the lexicon and grammatical contexts for geographical named entities. In the process of building a model of medieval language and a specific lexicon, the available tool can already be used to annotate texts and shows a significant improvement when compared with previous modules. However, most work remains to be done in terms of adding specific gazetteers for entities other than geographical.

Keywords

geographical named entities, NERC, place names

1 Introdução

O reconhecimento automático de topónimos foi atendido nas últimas duas décadas como parte do problema NERC (Named Entity Recognition and Classification) também chamado de REM (Reconhecimento de Entidades Mencionadas) dentro do Processamento da Linguagem Natural (PLN). Dado que os topónimos mais comuns aparecem sistematizados em nomenclatores e atlas já digitalizados, o uso de listas de entidades (*gazetteers*) sobre os que efetuar pesquisas por *string match* aparece como uma primeira solução para a anotação automática. Porém, a ambiguidade que se produz na língua (ex. *Santiago* como cidade ou como nome de pessoa) e a necessidade

DOI: 10.21814/lm.11.1.291
This work is Licensed under a Creative Commons Attribution 4.0 License

de marcar topónimos menores ou menos comuns (ex. microtopónimos, geografias exóticas e menos habituais) precisa de utilidades com capacidade de desambiguação e análise do contexto para prever os casos de formas desconhecidas nas listas. Duas aproximações contribuíram para dar o problema como resolvido com um nível satisfatório de eficácia: a aplicação de heurísticas (especialmente regras que especifiquem um contexto morfossintático) e o treino a partir de grandes volumes de corpora de onde se inferem regras de tipo estatístico. A anotação NERC passou assim a formar parte dos pacotes de utilidades de PLN mais comuns na atualidade, facilitando o processamento de textos para o reconhecimento de topónimos.

Dado que as ferramentas NERC foram desenvolvidas a partir de corpora contemporâneos (Won et al., 2018), a aplicação em variedades históricas da língua vê comprometido o desempenho em função da divergência a respeito dos usos linguísticos atuais. No caso galego-português medieval, ainda conservando uma estrutura gramatical e regularidade morfológica próxima às soluções contemporâneas, a dificuldade vem dada pelo grande número de variantes dos topónimos, limitando a aplicabilidade das listas (ex. as formas *Mondodnedo, Mondonedo, Mondonnedo, Mondoñedo* aparecem todas num mesmo *corpus*). O recurso a contextos sintáticos activados por palavras chave (*triggers*) que contribuem para a deteção da entidade geográfica com maior precisão, também se vê condicionado pelo fenómeno da variação (ex. *feegresia, figleſia, figressia, figresya, figrigia, figrisia...* até 438 variantes foram achadas para o mesmo tipo geográfico). Quanto menor seja a aplicabilidade de listas de entidades e de ativadores, mais limitação nos recursos da ferramenta e maior dependência na especificidade das regras ou no treino estatístico. Porém, para conseguir regras mais específicas, necessita-se um maior nível de PLN, particularmente lematização e etiquetagem morfossintática que, da sua parte, requer o uso de lexicons específicos para a variedade de língua. Numa solução estatística, o treino de modelos necessita de grandes corpora, com um volume suficiente como para serem estatisticamente relevantes. Para além de que este tipo de recursos são custosos e requerem atenção experta, existe ainda o problema de que a enorme produtividade da variação, acentuada por fatores tais qual época, área geográfica, tipologia textual e ainda usos individuais, faz com que as variantes se multipliquem e reduzam as frequências dos termos. Mesmo que um sistema NERC para textos medievais possa se desenhar com a mesma tecnologia e práticas utilizadas para a língua contemporânea, a adaptação duma ferramenta requer a disponibilidade de recursos adicionais que comprometem o desempenho do produto final. A nossa principal contribuição é a criação de recursos para o galego-português medieval e a sua integração e adaptação a uma ferramenta de NERC já existente para a língua contemporânea.

Mais precisamente, neste artigo apresentamos uma metodologia que analisa os componentes do pacote de utilidades PLN LinguaKit (Gamallo & Garcia, 2017) com maior relevância para a anotação de topónimos em textos medievais do domínio galego-português. Partindo dum conjunto de corpora com topónimos anotados manualmente, preparamos uma série de testes para avaliar o desempenho da ferramenta conforme se foram adicionando ou modificando componentes, nomeadamente listas de entidades e ativadores, até desenvolvermos um novo módulo NERC de LinguaKit adaptado a textos medievais do galego-português. Mesmo se este módulo é apenas uma primeira versão a melhorar, oferece um incremento muito notável na abrangência e medida-F para as entidades geográficas a respeito dos módulos de língua contemporânea. O módulo contém também um tokenizador, um lematizador e um etiquetador morfossintático, ainda em fase de protótipo, todos eles adaptados para o galego-português medieval. O conjunto de módulos para a língua histórica, chamado de *histgz*, está integrado em LinguaKit, com licença livre GPLv3[1].

Para além da introdução, o resto do artigo está organizado como prossegue. Na Secção 2 mencionamos estratégias de aproximação a textos históricos para o reconhecimento de entidades mencionadas e revemos o desempenho de ferramentas NERC para o caso particular de entidades geográficas mencionadas em português. A Secção 3 descreve os corpora utilizados nos testes e introduz LinguaKit, a ferramenta sobre a que se obtém a nova utilidade, cujas fases de desenvolvimento são atendidas na Secção 4. A seguir, avaliam-se os resultados, com atenção particular a falsos positivos e falsos negativos que apontam para melhoras mais imediatas e trabalho futuro, recolhido, junto com as conclusões, na Secção 6.

2 Trabalho relacionado

A dificuldade de reconhecer entidades geográficas mencionadas em textos antigos favorece a anotação manual, mais ou menos auxiliada por ambientes que facilitem o labor experto e possibilitem o trabalho em equipa. Na procura de maior

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

automatização, tem-se recorrido a sistemas inicialmente concebidos para o processamento de textos contemporâneos, com duas linhas de atuação, seja pela adaptação dos recursos utilizados pela ferramenta, particularmente listas de entidades, ou adaptando o texto, aplicando estratégias de normalização com o objetivo de minimizar a variação linguística e aproximar a língua histórica ao padrão contemporâneo (Hendrickx & Marquilhas, 2011; Marquilhas & Hendrickx, 2014)

As métricas mais comuns para a avaliação de ferramentas NERC são a precisão, a abrangência e a medida-F que combina as duas primeiras (Santos et al., 2007; Pinto et al., 2016). Na década passada celebraram-se eventos em que corpora previamente anotados serviam de padrões dourados para medir o desempenho de utilidades NERC. Especialmente relevantes para o português contemporâneo foram as competições do HAREM (Santos & Cardoso, 2007; Mota & Santos, 2008; Freitas et al., 2010). Os melhores resultados para a categoria de Lugar situaram-se em 68,03% de precisão e 73% de abrangência no primeiro evento (Santos & Cardoso, 2007) e precisão 72,12%, abrangência 80,17% no segundo (Chaves, 2008).

Ainda no domínio do padrão contemporâneo, testado no corpus Bosque (Afonso et al., 2002), o sistema NERC que deu lugar aos módulos correspondentes de LinguaKit atingiu 85% de precisão e 57% de abrangência (68% medida-F) na categoria de Lugar (Gamallo & Garcia, 2011). Este mesmo sistema foi também adaptado para a variedade linguística galega, com resultados de medida-F de entre 74,5% e 80,4%, em função do tipo de avaliação realizada (Garcia et al., 2012).

Testes mais recentes, utilizando os mesmos corpora que o HAREM, ofereceram resultados mais baixos, de 62% precisão e 66% abrangência para a mesma categoria de Lugar (Amaral et al., 2014), o qual é indicativo de que não houve um avanço significativo na resolução do problema.

Em relação com o reconhecimento de entidades em textos antigos, a maior parte dos trabalhos têm implementado estratégias determinísticas que combinam listas de entidades com heurísticas para cada tipo de entidade. Tanto as caraterísticas deste tipo de documentos, muitas vezes digitalizados mediante OCR, como o seu tamanho fazem com que a implementação de sistemas estatísticos seja mais custosa. Existem, contudo, sistemas baseados em aprendizagem automática, tais como Byrne (2007), que treina um modelo de máxima entropia orientado principalmente à identificação de entidades que se sobrepõem.

Dentro dos métodos determinísticos as aproximações mais frequentes são centradas no documento (i.e., utiliza-se informação de todo o texto para classificar uma entidade, e não só o contexto da menção específica a analisar). Assim, Jones & Crane (2006) classificam 10 tipos de entidades num jornal americano do século XIX, com valores de precisão de entre 57% e 99% em função da classe. Borin et al. (2007) analisam as entidades mencionadas num corpus de literatura sueca do XIX, melhorando os resultados com um módulo de similaridade que computa a distância de edição entre as menções desconhecidas e as listas de entidades. Também mediante máquinas de estados finitos e um conjunto de listas (nomes, apelidos, etc.), Grover et al. (2008) identificam entidades geográficas e nomes de pessoa em textos parlamentares britânicos dos séculos XVII a XIX.

De modo similar a estes últimos, a metodologia empregada no presente trabalho adapta os conjuntos de ativadores e de listas de entidades e implementa regras específicas para melhorar o desempenho dum sistema NERC em texto medieval.

Ainda que o número de trabalhos de anotação de topónimos aumenta particularmente no campo das humanidades digitais, são poucos ainda os estudos específicos para a comparação do desempenho de ferramentas NERC em textos históricos. Canosa (2017) comparou o desempenho de ferramentas para um corpus em inglês do século XVII atingindo resultados do 68% na medida-F com uma ferramenta estatística treinada em corpora modernos e do 62% com um sistema de regras provisto duma lista específica. Resultados similares, com a melhor medida-F próxima a 70%, foram de novo obtidos na comparativa de cinco ferramentas NERC contemporâneas sobre córpora históricos também do inglês em Won et al. (2018). Estes mesmos autores apresentam um experimento novidoso em que se combinam as anotações das distintas ferramentas NERC para escolher o mais provável em caso de divergência, atingindo um 73,3% na medida-F como melhor resultado.

3 Materiais e ferramentas

Nesta secção, apresentamos os recursos textuais (corpora) e a ferramenta de processamento da língua natural utilizados na experimentação (LinguaKit).

3.1 Corpora

Como material de trabalho para o treino duma nova ferramenta e avaliação de resultados utilizou-se uma parte dos textos medievais recolhidos actualmente no *Corpus informatizado Galego-Português Antigo*[2] (CGPA) (Varela Barreiro et al., 2016).

Código corpus	Tokens	Topónimos
Mens	18711	381
Toxosoutos	44001	2945
Toxosoutos_gl	5138	295
CDMACM5	267965	2626
CDMACM5_gl	105172	844
Mens_TX_CD_gl	128992	5760

Tabela 1: Códigos dos *corpora* utilizados para extração de topónimos e tamanho dos corpora e listas obtidas.

O CGPA é o resultado de reunir numa plataforma conjunta corpora históricos do galego, do português, do latim e do castelhano elaborados na Galiza, Portugal e Brasil. O núcleo de textos da Galiza forma-o o corpus plurilíngue *Xelmírez, Corpus lingüístico da Galiza medieval*[3] (Varela Barreiro, 2009) do Instituto da Lingua Galega (USC), em que estão integrados textos redigidos em galego-português (TMILG), em latim (TMILL) e em castelhano (TMILC). Os textos de Portugal e do Brasil não contêm obras em latim ou castelhano e estão representados por duas vias. Por parte portuguesa concorre o *Corpus Informatizado do Português Medieval*[4] (CIPM) (Xavier, 2000) e por parte brasileira o *Corpus Histórico do Português Tycho Brahe*[5] (Galves, 2018). Pelo momento o grande valor do CGPA é fazer possível, por meio de pesquisa única, o acesso à totalidade dos corpora integrados.

Os textos do CGPA selecionados para este projecto particular de desenvolvimento duma ferramenta de anotação de entidades geográficas mencionadas têm a particularidade de contarem com a etiquetagem dos topónimos, por quanto foram utilizados anteriormente no *Inventario Toponímico da Galiza Medieval*[6] (ITGM) (Varela Barreiro & Martínez Lema, 2009). O ITGM é um projecto lançado em 2005 com o intuito de fazer acessível de forma gradual a totalidade do material toponímico presente na documentação galega medieval, compilada e codificada no corpus *Xelmírez*. No processo de recuperação da informação, as agrupações de topónimos obtiveram-se pela aplicação de critérios linguísticos (relativos fundamentalmente ao processo de lematização) e/ou de critérios geográfico-administrativos. No seu estado actual, o ITGM acolhe 17.640 registos toponímicos, que remitem a um total de 3.086 topónimos e outros tantos lemas. Destes últimos, 2.876 (93% do conjunto) estão referenciados com maior ou menor margem de certeza, no entanto apenas para 7% (os 210 restantes) carecemos de qualquer parâmetro geográfico-administrativo de atribuição.

A tabela 1 mostra os textos procedentes do ITGM, caracterizados portanto por terem os topónimos anotados manualmente, que foram selecionados para a fase de recolhida de dados e testes no presente projecto. Cada corpus recolhe textos medievais com uma mesma origem documental, acessível no CGPA em que aparecem com código e referência bibliográfica individual. As obras do CGPA escolhidas para o desenvolvimento foram a *Colección diplomática do mosteiro de Santiago de Mens* (Zapico Barbeito, 2005), *Os documentos do tombo de Toxos Outos* (Rodríguez & Javier, 2004) e a *Colección diplomática medieval do Arquivo da Catedral de Mondoñedo* (Cal Pardo, 1999). Dado que nos corpora iniciais há também documentos em latim e castelhano, gerou-se uma segunda versão só com os textos em galego-português (identificada com a extensão *gl*). Nos experimentos finais agrupam-se todos os documentos num único arquivo (MensTXCDgl), a soma de selecionar só os documentos galego-portugueses dos três corpora. Na fase final de avaliação utilizou-se um novo corpus, o *Livro de Notas de Álvaro Pérez* (LNAP) (Tato Plaza, 1999), sem anotação de topónimos nenhuns na versão oferecida para este projeto.

3.2 LinguaKit

LinguaKit é um pacote livre de ferramentas multilíngues para o processamento da linguagem natural que pode aplicar-se ao português, galego, inglês e espanhol. Contém módulos de análise, extração, anotação e correção linguística. LinguaKit permite realizar um amplo conjunto de tarefas, entre as quais se encontram: segmentação em frases e tokenização, lematização, etiquetagem morfossintática, reconhecimento e classificação de entidades mencionadas (NERC), análise sintática de dependências, resolução de correferência a nível de entidade, extração de termos e de relações semânticas, análise de senti-

[2] http://ilg.usc.gal/CGPA
[3] http://sli.uvigo.gal/xelmirez/
[4] https://cipm.fcsh.unl.pt/
[5] http://www.tycho.iel.unicamp.br/~tycho/corpus/
[6] http://ilg.usc.gal/itgm

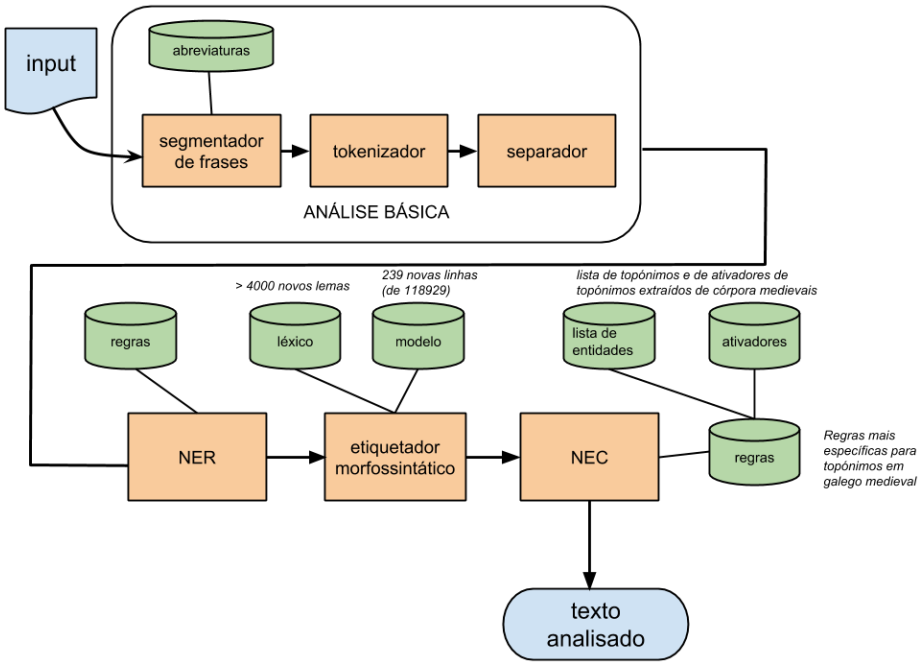

Figura 1: Arquitetura utilizada pelo Histgz e principais modificações a respeito dos módulos já existentes no LinguaKit.

mentos, anotação conceitual com ligação a recursos enciclopédicos (*entity linking*), correção e avaliação de léxico e sintaxe (só para o galego), conjugação verbal automática (exceto inglês), resumo automático, identificação de língua e visualização de concordâncias (palavras chave em contexto). O presente trabalho foca-se nas tarefas relacionadas com o NERC, tal e como mostra a figura 1, descrita mais à frente, na secção 4.

O LinguaKit está disponível como um serviço web[7] e é acessível via RESTful API4.[8] O código fonte está publicado sob licença GPLv3 e acessível desde repositório de GitHub (Cf. nota de rodapé 1).

4 Métodos e procedimento

O trabalho de adaptação do LinguaKit para o reconhecimento de topónimos medievais realizou-se aplicando uma metodologia experimental. Considerado um parámetro, aplicaram-se testes sobre os corpora para observar como contribui para a anotação dos topónimos. Primeiramente atendeu-se à incidência da lista de topónimos e a lista de ativadores. As regras e ajustes menores nos recursos de tipo morfossintático ocuparam o estádio final. Os corpora foram introduzidos no desenvolvimento gradualmente, de tal modo que uma vez concluída a necessidade de incluir um componente de melhora no módulo, se adicionava um novo corpus para a nova rolda de experimentos. As adaptações, ajustes, modificações e suplementos elaborados nos sucessivos ensaios foram implementados na arquitetura do LinguaKit como módulo independente, chamado de Histgz (galego-português histórico). A figura 1 mostra graficamente as principais modificações do novo módulo a respeito dos módulos prévios utilizados como base para o desenvolvimento. O Histgz inclui tanto tarefas básicas de análise (segmentação em frases, tokenização e quebra de contrações ou *separador*), quanto processos mais complexos: etiquetação morfossintática (*PoS tagging*) e NERC. As principais modificações foram realizadas no NEC (que forma parte do NER) e no etiquetador. O NEC é a tarefa final que classifica as entidades e permite reconhecer os topónimos.

4.1 Extração e elaboração de listas de topónimos

Para facilitar os labores de validação das anotações obtidas por procedimentos automáticos, criaram-se duas versões dos corpora: uma só texto, com os topónimos sem anotar, para ser usada como input pela ferramenta NERC, e outra com os topónimos marcados segundo a anotação manual facilitada pela

[7] https://www.linguakit.com

[8] https://market.mashape.com/linguakit/linguakit-natural-language-processing-in-the-cloud

Lista de topónimos	Sem lista medieval	Sem lista medieval	lista acrescentada com Lista_Mens	Lista acrescentada com Lista_Mens
Entidade geográfica mencionada	Verdadeiro positivo se coincide exatamente com a anotação manual	Verdadeiros positivos a partir do nome próprio	Verdadeiro positivo se coincide exatamente com a anotação manual	Verdadeiros positivos a partir do nome próprio
Precisão	57,53%	76,88%	72,97%	93,24%
Abrangência	17,95%	23,99%	45,3%	57,89%
Medida-F	27,36%	36,57%	55,9%	71,43%

Tabela 2: Comparativa de resultados da anotação do LinguaKit (com o módulo base em galego) sobre o corpus Mens segundo lista e critério de validação dos verdadeiros positivos das entidades geográficas mencionadas.

equipa do ITGM que servirá de padrão dourado e recurso para a extração de topónimos e listas para adicionar ao LinguaKit. As listas de topónimos obtidas relacionam-se na última coluna da tabela 1.

A lista última é a soma das listas de topónimos obtidas dos corpora envolvidos no desevolvimento do módulo NERC medieval, Lista_Mens_TX_CD (tabela 1) e aparece integrada na lista de entidades geográficas do Histgz acessível no repositório de GitHub (Cf. nota 1). Os testes realizados nas fases de desenvolvimento utilizaram as listas progressivamente, para distinguir o efeito da inclusão de cada nova lista segundo se processava e ensaiava sobre um novo corpus. Este efeito mostra-se na tabela 2 com os resultados do LinguaKit em galego condicionados pelo uso ou não da lista de entidades geográficas. As métricas mostram o limite de capacidade do módulo quando a lista contém todos os topónimos presentes no texto. Na avaliação dos resultados apareceu também como relevante o critério utilizado para definir o topónimo. O padrão dourado anota como topónimos frases do tipo "o rrío de Tallo", "no Esto" ou o artigo mesmo quando não foi grafado com maiúscula "o Esto", porém a lista de entidades limita o topónimo ao nome próprio. Nas primeiras avaliações discriminou-se entre os resultados que coincidiam plenamente com as anotações manuais (só se avalia como verdadeiro positivo quando se anota exatamente igual ao padrão, assim "o rrío de Tallo", "no Esto") face àqueles em que o nome próprio é suficiente para considerar a anotação como verdadeiro positivo ("Tallo", "Esto"). Este último critério, mais adequado às expectativas reais de desempenho duma ferramenta NERC, será o que se aplique nos sucessivos experimentos.

Os experimentos da tabela 2 mostram como, mesmo com uma lista que contém todos os topónimos presentes no corpus, ainda obtendo uma precisão muito alta, apenas se recuperam 57,89% das entidades geográficas mencionadas. As regras utilizadas pelos módulos de língua contemporânea precisam, portanto, melhoras para além das listas que, contudo, resultam determinantes para um bom desempenho (os resultados com a lista só de topónimos contemporâneos ficam em apenas 36,57% na medida-F face ao 71,43% obtido ao acrescentar a lista medieval).

4.2 Filtrado por língua

A análise dos primeiros testes, uma vez processados os textos e extraídos os topónimos anotados manualmente, mostraram as limitações da aplicação do módulo NERC mesmo com uma lista de topónimos específica. Porém, o primeiro factor a considerar para a melhora de resultados não é devido ao pacote PLN, mas aos próprios textos a anotar. Com efeito, ao trabalhar com o conjunto dos corpora, aparecem textos em latim e, em menor medida e em documentos mais tardios, espanhol, que necessariamente condicionam a efetivadade dos recursos e das heurísticas classificatórias, dependentes duma seleção linguística prévia (ex. *illa* é demonstrativo em latim, mas nome comum "terra rodeada por mar" em galego). Faz-se necessária, portanto, a discriminação por idioma. Os corpora utilizados nos vindouros experimentos (Toxosoutos e CDMACM5) foram processados para obter apenas o texto galego(-português), identificado com a extensão _gl nos códigos da tabela 1.

Precisão	**Abrangência**	**Medida-F**	**lista com os topónimos do corpus**	**Lista de ativadores**
82,26%	25,67%	39,13%	Não	Não
96,22	59,73%	73,71%	Sim	Não
77,5%	36,41%	49,54%	Não	Sim
90,31%	62,58%	73,93%	Sim	Sim

Tabela 3: Comparativa de resultados segundo o uso de lista de ativadores e topónimos no *corpus* Mens (596 entidades geográficas anotadas no texto padrão).

Ativadores	Lista de ativadores expandida por combinatória de caracteres	Lista de ativadores recuperada por inspeção de concordâncias no CGPA
Precisão	76,74%	76,16%
Abrangência	32,76%	33,06%
Medida-F	45,92%	46,11%

Tabela 4: Comparativa de resultados variando a lista de ativadores sobre o corpus Mens_TX_CDMACM5_gl (3.373 entidades geográficas anotadas no corpus padrão).

4.3 Lista de ativadores

O primeiro componente considerado para a melhora do desempenho do módulo uma vez comprovada a limitação da lista de topónimos foi a lista de ativadores para entidades geográficas (TwLOC). Da inspecção experta de concordâncias dos topónimos extraídos do primeiro corpus usado nos testes (Mens) obteve-se manualmente uma lista de 38 termos geográficos, contando como unidades distintas todas as variantes dum mesmo termo. Assim, na mesma lista aparecem todas as variantes achadas no *corpus* associáveis a *fregresia* (*fregesía, fijglesía, fijgresía, fjgresía, flegresía, frijguesía, frigresýas*), mosteiro (*moesteiro, moesteyro, mosteiro, mosteiros*), vila (*vila, villa, vjla*) junto com termos geográficos com uma única ocorrência (*tença*).

Dado que uma boa parte dos termos são variantes do mesmo tipo, experimentou-se com a expansão da lista de ativadores mediante a combinatória de caracteres equivalentes (ex. nasal palatal *nn*, *nh*, *nj*, *ni*, *jn*, *yn*, *in*, *ñ*, *gn*; vogais simples e geminadas *o*, *oo*; vogais nasais e terminações *õ*, *om*, *on*; sibilantes, lateral palatal, etc.).

Paralelamente fez-se um levantamento manual de termos geográficos a partir das listas de tipos extraídas do CGPA de onde se obtêm 1.900 termos geográficos (disponíveis na pasta de ativadores do próprio módulo Histgz de LinguaKit).

Os resultados da aplicação da lista de termos expandidos artificialmente não proporcionaram um incremento sobre a lista manual (tabelas 3 e 4) e, toda vez que esta contém as formas originais dos corpora, ficou esta última como a solução finalmente adoptada para o novo módulo.

Ao tempo que se elaborou a nova lista de ativadores com termos geográficos, recolheram-se termos adicionais com valor de ativador em contextos mais específicos, agrupados em uma lista, *nongeo*, composta principalmente por títulos, ex. *arcebispo*, *rei*. Recolhe 686 termos, com um alto número de variantes para o mesmo tipo.

4.4 Regras classificatórias

O módulo NERC aplica a lista de ativadores por meio de regras que priorizam a classificação numa classe dentro das quatro categorias de entidades mencionadas (PER Pessoa, LOC Lugar, ORG Organização e MISC Miscelânea). Um exemplo de regra é aquela que classifica um nome próprio como entidade geográfica mencionada quando não se achar em nenhuma das listas de entidades e o termo precedente for a preposição *em*. Outra classifica como nome de pessoa todo nome próprio ambíguo em ausência de outros condicionantes. Assim, se uma expressão aparece em ambas as listas de pessoas e topónimos, será considerada PER e não LOC por quanto os antropónimos são mais frequentes do que os nomes de lugar. Resulta óbvio que as regras têm um rendimento percentual e não representam o 100% dos casos (ex. *Penso em Ruy* daria erro com a primeira regra citada da preposição *em*, e *Santiago* seria classificado sempre como nome de pessoa se estiver em ambas as duas listas de entidades LOC e PER e não houvesse contexto nenhum para a desambiguação).

Regras complementares, como a aplicação de contextos gramaticais e as listas de ativadores, permitem corrigir parcialmente os erros derivados das regras mais genéricas. Como o problema é classificatório, quanto melhor se discriminem as outras categorias, melhores resultados se obterão na classe de entidades geográficas. Porém, dado que nesta adaptação de LinguaKit o foco de estudo foram os topónimos, as regras características do módulo consideram exclusivamente as entidades geográficas, desambiguando os contextos gramaticais e precisando formas específicas da língua medieval (caso das contrações da preposição *em,* ex. *enno*). O recurso a listas de categorias não geográficas faz-se quando é possível criar uma regra que melhore a recuperação de topónimos. Assim uma lista de ativadores para entidades da categoria pessoa (ex. *bispo*, *emperatriz*, *rrainha*, *rei*, etc.), extraída ao tempo que se recolheram os ativadores *geo*, permite classificar como nomes de pessoa os nomes próprios precedidos dos termos que assinalam a entidade pessoa (PER) em casos como:

bispo <PER>*Payo Rodrigues*</PER>

No entanto, ante preposição *em* ou *de* e, opcionalmente, artigo, as regras aplicam a lista para reconhecer uma entidade geográfica:

bispo de <LOC>*Mondonedo*</LOC>

Idealmente, o sistema devia também reconhecer e classificar *bispo de Mondonedo* como pessoa mas o trabalho presente está focado no reconhecimento de topónimos e a ferramenta de extração não foi configurada para identificar entidades dentro da expressão duma outra entidade.

4.5 Lexicon e modelo de língua

Para a etiquetagem morfossintática prévia à classificação das entidades mencionadas, LinguaKit usa um lexicon que regista o lema ou lemas e valores gramaticais de cada expressão. Na aussência dum corpus lematizado e etiquetado morfossintamente que permitisse capturar recursos e treinar um modelo específico de língua medieval, juntaram-se os lexicons de galego e português acrescentados com um novo dicionário criado a partir dos termos de maior frequência (mínimo 20 ocorrências) nos corpora usados na fase de desenvolvimento (tabela 1). Para a desambiguação da etiqueta morfossintática (ex. *era*: nome comum feminino singular ou verbo imperfeito indicativo da primeira ou terceira pessoa), utilizou-se o modelo de galego, preferido por quanto mantém os pronomes enclíticos ligados ao verbo sem marca nenhuma. Parte dos textos medievais anotados automaticamente nos testes mencionados nas secções anteriores foram revistos para serem utilizados num treino mínimo de bigramas de tokens adicionado ao modelo contemporâneo. O modelo criado é utilizado por um desambiguador bayesiano para levar a cabo a etiquetagem morfossintática tal como descreve Garcia & Gamallo (2015).Mais concretamente, este módulo é um classificador bayesiano baseado em bigramas de pares $< token, etiqueta >$. Para poder atribuir uma marca (ou etiqueta morfossintática) a um token, o classificador calcula a probabilidade de cada marca dado o token alvo tomando em conta o contexto à esquerda e à direita, nomeadamente tomando em conta as etiquetas imediatamente à esquerda e à direira do token alvo. O algoritmo desambigua de esquerda à direita, de tal maneira que o contexto esquerdo dum token ambíguo é um outro token já desambiguado, é dizer, ao qual já foi atribuído uma única etiqueta.

4.6 Ajustes finais

A aplicação do módulo sobre os corpora de desenvolvimento representa uma aproximação ao que seria o limite máximo de anotação do novo módulo quando operar nas condições idóneas de abrangência total da lista de entidades geográficas e textos com topónimos cujos tipos geográficos aparecem recolhidos na lista de ativadores. A tabela 5 mostra o resultado de variar estes componentes com as regras atualizadas. Nas melhores condições, mesmo com uma lista que contém todos os topónimos do corpus a anotar, a abrangência máxima apenas atinge o 60%.

Os últimos testes serviram para confirmar a configuração ótima a respeito dos parâmetros em que a diferença no desempenho resultou ser menor (expansão da lista de ativadores e regras específicas) e realizar correções menores e pontuais em casos de ambiguidades que, sendo muito específicas dos documentos utilizados no treino, podiam afetar negativamente o desempenho noutros textos.

5 Avaliação

Para a avaliação da configuração final do módulo Histgz utilizou-se o *corpus* LNAP, presente na coleção do CGPA, mas não utilizado nas fases de desenvolvimento. O texto também não tem nenhuma anotação de topónimos e apenas requereu um pré-processamento básico para ser enviado como input para o LinguaKit. Para a validação dos verdadeiros e falsos positivos e nega-

Topónimos	Sem lista de topónimos	Com lista de topónimos
Ativadores	Sem lista de ativadores	Com lista de ativadores
Precisão	86,15%	85,12%
Abrangência	14,2%	60,36%
Medida-F	24,38%	70,63%

Tabela 5: Configurações finais do Histgz para a anotação dos próprios corpora usados no treino Mens_TX_CDMACM5_gl (3.373 entidades geográficas anotadas no padrão)

FALSOS NEGATIVOS
morador morador NCMS000 ẽna en+a SPS00+DA freigresja fregresia NCFS000 de de SPS00 **Santa_Coõba_de_Rriãjo santa_coõba_de_rriãjo NP00SP0**
En en SPS00 **San_Tomé_de_o_Mar san_tomé_de_o_mar NP00SP0**
morador morador NCMS000 ẽna en+a SPS00+DA dita dita NCFS000 **Sã sã NP00O00** **Mjgell mjgell NP00V00**

Tabela 6: Exemplos de avaliações com falsos negativos sobre a anotação do Histgz. As etiquetas NP00SP0, NP00O00, NP00V00 representam respetivamente as classes PER(soa), ORG(anização) e MISC(eláneo).

tivos (tabela 6) reviram-se todos os outputs manualmente. Dado que este é um trabalho custoso, utilizaram-se apenas os documentos iniciais do corpus até superar os 2.000 tokens (12 documentos com 2.060 tokens em total). A tabela 6 oferece uma pequena amostra da saída de Histgz, onde cada unidade lexical do texto de entrada se divide em três colunas: *token*, lema e etiqueta. O conjunto de etiquetas empregado tem por volta de 250 etiquetas mofossintáticas e baseia-se nas recomendações do Grupo EAGLE.[9]

5.1 Resultados

Visto que não há ferramentas específicas para trabalho com língua medieval, a avaliação de resultados centra-se na melhora do novo módulo, Histgz, com os já existentes para língua contemporânea no próprio LinguaKit. O gráfico da figura 2 mostra os resultados obtidos pela validação manual das anotações obtidas com as configurações dos módulos Histgz (galego-português medieval), gl (galego) e pt (português) sobre o texto com os documentos do corpus LNAP. O teste mostra o incremento do desempenho a respeito dos módulos já existentes no pacote do LinguaKit para galego e português atuais (figura 2). A abrangência, a medida que mais se vê condicionada pela adequação da lista de topónimos ao corpus (cf. tabelas 3, 4 e 5), é a que mostra um incremento mais notável, ao nível mesmo da obtida nos testes mais favoráveis durante a fase de desenvolvimento (cf. tabela 5). A precisão também obtém um melhor rendimento a respeito das versões de língua contemporânea, porém o desempenho é menor que o obtido nos testes de desenvolvimento, mesmo nas situações mais adversas (cf. tabela 3). A comparação com outros sistemas fica fora dos objetivos presentes, por quanto o Histgz é um produto operativo mas em estado muito inicial e por serem os testes realizados sobre um texto ainda que não utilizado no treino, sim presente na mesma coleção do CGPA a que pertencem os corpora usados na fase de desenvolvimento. Apenas como referência das expectativas que se podem aguardar dum sistema de reconhecimento distinto a

[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html

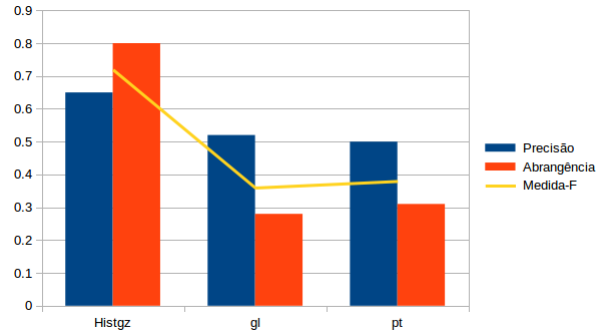

Figura 2: Comparativa do desempenho das configurações por língua do LinguaKit para um texto medieval não utilizado na fase de desenvolvimento. Os módulos *gl* e *pt* representam o Galego e Português contemporâneos, respectivamente.

LinguaKit, realizou-se um teste com os módulos de Português e Galego modernos de Freeling 4.1 sobre o mesmo texto de teste com o que foi avaliado o módulo Histgz. Os valores de medida-F obtidos foram de 44% (módulo do Português) e 39% (módulo de Galego). Estes resultados mostram um desempenho maior do que os módulos de Português e Galego modernos de LinguaKit, porém bem abaixo do nosso módulo histórico, Histgz, que, tal e como se observa na Figura 2, consegue uma medida-F superior ao 70%.

5.2 Discussão a análise de erros

A inspecção dos falsos positivos permite apontar melhoras imediatas pelo afinamento das regras e priorizar atuações nos componentes do Histgz.

5.2.1 Falsos negativos

Assim, como exemplo de falsos negativos corrigíveis por regras, um topónimo precedido por um termo geográfico (*freigresia*) é etiquetado como entidade pessoa por quanto contém também uma expressão que é internamente lematizada como ativador da classe PER (*santo*), que o classificador prioriza frente à classe LOC em casos de ambiguidade (por ex: *Santa_Coõba_de_Rriãjo*, falso negativo, tabela 6). Mesmo em contextos em que o classificador favorece a entidade geográfica, quando a ambiguidade se produz depois da preposição *em*, a presença dum token reconhecido como nome de pessoa (PER) na lista de entidades, faz com que o classificador dirima em favor desta classe (*San_Tomé_de_o_Mar*, falso negativo, tabela 6). Quando a entidade não é recuperável nem total nem parcialmente nas listas de ativadores ou entidades, a anotação já se vê afetada pela segmentação pois nenhum dos termos é reconhecido (*Sã Mjgell*, falso negativo, tabela 6).

A grande presença de topónimos hagiográficos nos textos medievais aconselha uma maior especificação na aplicação desta regra de desambiguação.

5.2.2 Falsos positivos

No caso dos falsos positivos, a anotação dos numerais romanos como nomes próprios é uma mostra da necessidade de melhorar o lexicon, factível de modo mais imediato neste caso dos numerais, frequentes e sistematizáveis (*XXVJ*, falso positivo, tabela 7). A carência duma lista de nomes de pessoa medievais provoca a classificação como entidade geográfica dos antropónimos (*Rroý_Bouçón*, falso positivo, tabela 7) quando aparecem perto de termos geográficos (uso da lista de ativadores ao não ser reconhecido o termo na lista de entidades). Dado que os corpora de desenvolvimento anotam também os antropónimos, a elaboração duma lista de nomes de pessoa medieval é também suscetível de melhora imediata.

5.2.3 Critérios para a consideração dos topónimos como verdadeiros ou falsos positivos

A análise dos falsos positivos e negativos durante a validação dos resultados mostrou que as métricas de desempenho vêm muito condicionadas pelos critérios utilizados para a definição da entidade geográfica mencionada. Nos corpora usados na fase de desenvolvimento, nomes próprios que seguem um antropónimo aparecem ocasionalmente anotados como topónimos e, muito mais frequentemente, quando precedidos pela preposição *de*. Assim, em casos como *Domingo_**Vidal*** e *Diego_Sanches_de_**Ribadeneyra***, considerados em âmbitos NERC como entidade mencionada de pessoa no seu conjunto, as formas em negrito vieram anotadas como topónimos nos corpora de treino e, consequentemente, foram avaliados como falsos negativos caso de não serem reconhecidos como geográficos nos resultados do LinguaKit. Porém, o critério utilizado para a elaboração das regras do Histgz utiliza a definição mais standard dos sistemas NERC, comum com os módulos das distintas línguas contemporâneas já presentes no pacote, de considerar uma única entidade mencionada multipalavra, mesmo se um dos elementos for também em origem classificável como entidade pertencente a outra classe. A aplicação dum critério que maximize o reconhecimento de topónimos indepen-

FALSOS POSITIVOS
dorna dorna NCFS000 de de SPS00 **XXVJ xxvj NP00G00** canadas canada NCFP000
morador morador NCMS000 ẽno en+o SPS00+DA0MS0 dito dito AQ0MS0 porto porto NCMS000 , , Fc a o DA0FS0 **Rroý_Bouçón rroý_bouçón NP00G00** , , Fc morador morador NCMS000 ẽna en+a SPS00+DA freigresja fregresia NCFS000
Testigos testigos NP00V00 : : Fd Vasco vasco AQ0MS0 de de SPS00 **Lees lees NP00G00**

Tabela 7: Exemplos de avaliações com falsos positivos sobre a anotação do Histgz. A etiqueta NP00G00 representa a classe LOC(alização) ou topónimo.

dentemente de qual for o referente principal ou, pela contra, a minimização do número de entidades em favor dum único referente, é discutível e varia em função dos interesses particulares da anotação. Em efeito, e mais particularmente nos textos medievais, uma mesma estrutura sintática pode ter tanto valor duma única entidade mencionada (<PER>*Vasco de Lees*</PER>) quanto de duas (<PER>*Vasco*</PER> *de* <LOC>*Lees*</LOC>). No caso da validação da anotação do Histgz, com o fim de aplicar um mesmo critério para todos os casos, entende-se que o nome próprio precedido por antropónimo mais preposição *de* deve ser validado como entidade mencionada de pessoa em todos os casos. Consequentemente, a anotação deste exemplo na tabela 7 foi avaliada como falso positivo. Esta divergência no critério do que é ou não uma entidade geográfica mencionada influi, portanto, nas métricas, e deve ser tida em conta à hora de valorar os resultados.

6 Conclusões e trabalho futuro

A adaptação para o trabalho com textos medievais duma ferramenta NERC inicialmente concebida para labores PLN com textos contemporâneos ofereceu uns resultados que melhoram notavelmente o desempenho a respeito dos módulos existentes. O labor de configuração consistiu na avaliação do desempenho da utilidade sobre corpora previamente anotados, modificando parâmetros e adicionando recursos segundo se ia experimentado com os textos. A principal dificuldade para a melhora dos resultados na medida-F é o incremento da abrangência sem comprometer excessivamente a precisão. A lista de topónimos foi considerada o recurso mais determinante, porém, dada a alta variação gráfica e a tipologia textual, que favorece a aparição de microtopónimos, qualquer lista moderna aparece comprometida na abrangência. A aplicação duma lista com termos geográficos permitiu melhorar os resultados, ainda que em menor grau do que a lista de topónimos, contudo, os ativadores contribuem também para a desambiguação de entidades de mais difícil classificação. O critério para a definição do que é ou não um topónimo e quando se deve reconhecer como entidade geográfica mencionada, influi na definição de regras e na validação dos resultados. Com a necessidade de salientar que a avaliação vem condicionada pela consideração de entidade geográfica mencionada mais conforme às práticas NERC do que a uma definição mais abrangente de topónimo, e que o corpus utilizado é reduzido para permitir uma validação manual, a análise do teste final mostra que o produto obtido melhora os resultados dos módulos NERC prévios. O Histgz é,

contudo, apenas uma versão inicial necessitada de melhoras. Ao atender preferentemente as entidades geográficas, ficaram desatendidos ou minimamente considerados outros recursos que contribuem tanto para o rendimento da anotação NERC quanto para a expansão das capacidades PLN dentro do amplo abano de utilidades que o LinguaKit oferece. O módulo no seu estado atual é maiormente dependente do treino sobre textos contemporâneos que serviram de base de desenvolvimento, com exceção da lista de ativadores e em menor medida da lista de entidades geográficas. O trabalho futuro consiste na ampliação do lexicon e o treino dum modelo de língua representativo da variedade medieval, labor que, pela dependência que tem na validação experta, requer de recursos notavelmente superiores aos utilizados para a obtenção do produto atual. Porém, o próprio módulo Histgz pode ser já aplicado para facilitar a preparação dos textos e produzir corpora que simplifiquem e agilizem o trabalho de reconhecimento e anotação.

Como já foi dito, o módulo foi integrado em LinguaKit e tanto o léxico como as listas de entidades e ativadores de Histgz estão disponíveis com licença livre.[10]

Agradecimentos

Este trabalho foi desenvolvido no marco da rede galega de investigação TECANDALI, ED341D R2016/011, financiada pela Consellaria de Educación e Ordenación Universitaria da Xunta de Galicia, e do European Regional Development Fund (ERDF).

Referências

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002. Floresta sintá(c)tica: a treebank for portuguese. Em *3rd International Conference on Language Resources and Evaluation (LREC)*, 1698–1703.

Amaral, Daniela, Evandro Fonseca, Lucelene Lopes & Renata Vieira. 2014. Comparative analysis of portuguese named entities recognition tools. Em *9th International Conference on Language Resources and Evaluation (LREC)*, 2554–2558.

Borin, Lars, Dimitrios Kokkinakis & Leif-Jöran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. Em *Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, 1–8.

Byrne, Kate. 2007. Nested named entity recognition in historical archive text. Em *International Conference on Semantic Computing (ISCS)*, 589–596.

Cal Pardo, Enrique (ed.). 1999. *Colección diplomática medieval do arquivo da Catedral de Mondoñedo. Transcrición íntegra dos documentos*. Santiago de Compostela: Consello da Cultura Galega.

Canosa, Afonso Xavier. 2017. *A identificacão e referenciacão de entidades geográficas mencionadas: o caso da 'Peregrinação' de Fernão Mendes Pinto*: Universidade de Santiago de Compostela. Tese de Doutoramento.

Chaves, Marcírio. 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no segundo HAREM. Em Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, Linguateca.

Freitas, Cláudia, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira & Paula Carvalho. 2010. Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. Em *7th International Conference on Language Resources and Evaluation (LREC)*, 3630–3637.

Galves, Charlotte. 2018. Tycho brahe parsed corpus of historical Portuguese. Universidade de Campinas. http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip.

Gamallo, Pablo & Marcos Garcia. 2011. A resource-based method for named entity extraction and classification. Em *Progress in Artificial Intelligence, 15th Portuguese Conference on Artificial Intelligence (EPIA)*, vol. 7026, 610–623. 10.1007/978-3-642-24769-9.

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

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. Em *Languages, Applications and Technologies (SLATE)*, vol. 563 Communications in Computer and Information Science, 65–75. 10.1007/978-3-319-27653-3_7.

[10] https://github.com/citiususc/Linguakit/tree/master/tagger/histgz

Garcia, Marcos, Iria Gayo & Isaac González López. 2012. Identificação e classificação de entidades mencionadas em galego. *Estudos de Lingüística Galega* 4. 13–25.

Grover, Claire, Sharon Givon, Richard Tobin & Julian Ball. 2008. Named entity recognition for digitised historical texts. Em *6th International Conference on Language Resources and Evaluation (LREC)*, 1343–1346.

Hendrickx, Iris & Rita Marquilhas. 2011. From old texts to modern spellings: An experiment in automatic normalisation. *Journal for Language Technology and Computational Linguistics* 26(2). 65–76.

Jones, Alison & Gregory Crane. 2006. The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection. Em *6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 31–40.

Marquilhas, Rita & Iris Hendrickx. 2014. Manuscripts and machines: the automatic replacement of spelling variants in a Portuguese historical corpus. *International Journal of Humanities and Arts Computing* 8(1). 65–80. 10.3366/ijhac.2014.0120.

Mota, Cristina & Diana Santos. 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo HAREM. Em Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, Linguateca.

Pinto, Alexandre, Hugo Gonçalo Oliveira & Ana Oliveira Alves. 2016. Comparing the performance of different NLP toolkits in formal and social media text. Em *5th Symposium on Languages, Applications and Technologies (SLATE)*, vol. 51, 3:1–3:16. 10.4230/OASIcs.SLATE.2016.3.

Rodríguez, Pérez & Francisco Javier (eds.). 2004. *Os documentos do tombo de Toxos Outos*. Santiago de Compostela: Consello da Cultura Galega.

Santos, Diana & Nuno Cardoso (eds.). 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área*. Linguateca.

Santos, Diana, Nuno Cardoso & Nuno Seco. 2007. Avaliação no HAREM: Métodos e medidas. Em Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 245–282. Linguateca.

Tato Plaza, Fernando R. (ed.). 1999. *Libro de notas de Álvaro Pérez, notario da Terra de Rianxo e Postmarcos*. Santiago de Compostela: Consello da Cultura Galega.

Varela Barreiro, Xavier (ed.). 2009. *Xelmírez. Corpus Lingüístico da Galiza Medieval*. Santiago de Compostela: Instituto da Lingua Galega (USC).

Varela Barreiro, Xavier & Paulo Martínez Lema. 2009. *Inventario Toponímico da Galiza Medieval*. Santiago de Compostela: Instituto da Lingua Galega.

Varela Barreiro, Xavier, Maria Francisca Xavier & Charlotte Galves. 2016. *Corpus informatizado Galego-Português Antigo*. Santiago de Compostela / Lisboa / Campinas: Instituto da Lingua Galega / Centro de Lingüística da Universidade Nova de Lisboa / Universidade de Campinas.

Won, Miguel, Patricia Murrieta-Flores & Bruno Martins. 2018. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities* 5. 2. 10.3389/fdigh.2018.00002.

Xavier, Maria Francisca (ed.). 2000. *Corpus Informatizado do Português Medieval*. Lisboa: Centro de Lingüística da Universidade Nova de Lisboa.

Zapico Barbeito, Pilar (ed.). 2005. *Colección diplomática do mosteiro de Santiago de Mens. Edición e estudo*. Noia: Toxosoutos.

Proposta recebida em Outubro 2018 e aceite para publicação em Julho 2019.

Reconhecimento de Actos de Diálogo Hierárquicos e Multi-Etiqueta em Dados em Espanhol

Hierarchical Multi-Label Dialog Act Recognition on Spanish Data

Eugénio Ribeiro
L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
Instituto Superior Técnico, Universidade de Lisboa, Portugal
eugenio.ribeiro@inesc-id.pt

Ricardo Ribeiro
L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
ricardo.ribeiro@inesc-id.pt

David Martins de Matos
L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
Instituto Superior Técnico, Universidade de Lisboa, Portugal
david.matos@inesc-id.pt

Resumo

Os actos de diálogo revelam a intenção por trás das palavras pronunciadas. Por isso, o seu reconhecimento automático é importante para um sistema de diálogo que tenta entender o seu interlocutor. O estudo apresentado neste artigo aborda essa tarefa no corpus DIHANA, cujo esquema de anotação de actos de diálogo em três níveis coloca problemas que não foram explorados em estudos recentes. Além do problema hierárquico, os dois níveis inferiores colocam problemas de classificação multi-etiqueta. Além disso, cada nível da hierarquia refere-se a um aspecto diferente relativo à intenção do orador, tanto em termos da estrutura do diálogo, como da tarefa. Por outro lado, uma vez que os diálogos são em espanhol, este corpus permite-nos avaliar se as melhores abordagens para dados em inglês generalizam para uma língua diferente. Mais especificamente, comparamos o desempenho de diferentes abordagens de representação de segmentos, com foco tanto em sequências como em padrões de palavras, e avaliamos a importância do histórico do diálogo e das relações entre os múltiplos níveis da hierarquia. No que diz respeito ao problema de classificação de etiqueta única colocado pelo nível superior, mostramos que as conclusões obtidas a partir de dados em inglês se mantêm em dados em espanhol. Para além disso, mostramos que as abordagens podem ser adaptadas para cenários multi-etiqueta. Por fim, combinando hierarquicamente os melhores classificadores para cada nível, obtemos os melhores resultados reportados para este corpus.

Palavras chave

reconhecimento de actos de diálogo, classificação hierárquica, classificação multi-etiqueta

Abstract

Dialog acts reveal the intention behind the uttered words. Thus, their automatic recognition is important for a dialog system trying to understand its conversational partner. The study presented in this article approaches that task on the DIHANA corpus, whose three-level dialog act annotation scheme poses problems which have not been explored in recent studies. In addition to the hierarchical problem, the two lower levels pose multi-label classification problems. Furthermore, each level in the hierarchy refers to a different aspect concerning the intention of the speaker both in terms of the structure of the dialog and the task. Also, since its dialogs are in Spanish, it allows us to assess whether the state-of-the-art approaches on English data generalize to a different language. More specifically, we compare the performance of different segment representation approaches focusing on both sequences and patterns of words and assess the importance of the dialog history and the relations between the multiple levels of the hierarchy. Concerning the single-label classification problem posed by the top level, we show that the conclusions drawn on English data also hold on Spanish data. Furthermore, we show that the approaches can be adapted to multi-label scenarios. Finally, by hierarchically combining the best classifiers for each level, we achieve the best results reported for this corpus.

Keywords

dialog act recognition, hierarchical classification, multi-label classification

DOI: 10.21814/lm.11.1.278
This work is Licensed under a Creative Commons Attribution 4.0 License

1 Introdução

Para um sistema de diálogo é relevante identificar a intenção por trás das palavras dos seus interlocutores, uma vez que esta fornece uma pista importante sobre a informação contida num segmento e como este deve ser interpretado. Segundo Searle (1969), essa intenção é revelada pelos actos de diálogo, que são as unidades mínimas de comunicação linguística. Consequentemente, o reconhecimento automático de actos de diálogo é uma tarefa importante no contexto da compreensão de língua natural, que tem sido amplamente explorada ao longo dos anos em múltiplos corpora com diferentes características. Recentemente, a maioria dos estudos tem-se focado em dados em inglês e, mais especificamente, no Switchboard Dialog Act Corpus (SwDA) (Jurafsky et al., 1997), uma vez que este é o maior corpus anotado com actos de diálogo e o seu conjunto de etiquetas é independente da tarefa e do domínio. No entanto, existem outros corpora e esquemas de anotação que colocam problemas no contexto do reconhecimento de actos de diálogo que não são cobertos pelo corpus SwDA e as suas anotações no formato SWBD-DAMSL. Tendo isso em conta, neste artigo exploramos o corpus DIHANA (Benedí et al., 2006), que contém interações em espanhol entre humanos e um sistema de diálogo simulado usando o método do Feiticeiro de Oz (WoZ). No contexto do reconhecimento de actos de diálogo, este corpus diferencia-se dos restantes pelo seu esquema de anotação em três níveis, no qual o nível superior se refere ao acto de diálogo genérico e independente da tarefa e os restantes o complementam com informação específica da tarefa. Para além disso, cada segmento tem apenas uma etiqueta de nível superior, mas pode não ter nenhuma ou ter várias etiquetas nos restantes níveis. Tendo em conta estas características, o corpus DIHANA permite-nos abordar o reconhecimento de actos de diálogo como um problema de classificação hierárquica e multi-etiqueta.

Similarmente ao que acontece com outras tarefas de classificação de texto, tais como categorização de notícias e análise de sentimento (Kim, 2014; Conneau et al., 2017), a maioria das abordagens recentes ao reconhecimento de actos de diálogo baseiam-se em Redes Neuronais Profundas (DNNs). Uma visão geral sobre essas abordagens é fornecida na Secção 2.2. No entanto, em geral, estas usam uma abordagem baseada em Redes Neuronais Recorrentes (RNNs) ou Redes Neuronais Convolucionais (CNNs) para gerar uma representação do segmento a partir da representação das suas palavras na forma de *embeddings*. Em seguida, a informação presente nessa representação é usada para obter a classificação do segmento. A distinção entre as abordagens baseadas em RNNs e CNNs é relevante, uma vez que estas são capazes de capturar diferentes tipos de informação. No caso das primeiras, o foco é em identificar sequências de palavras relevantes, incluindo dependências de longo alcance. Por outro lado, as últimas focam-se na identificação de padrões de palavras relevantes, observando janelas de contexto em torno de cada palavra. Para além disso, as abordagens com desempenho mais alto na tarefa não consideram cada segmento por si só, mas sim em conjunto com informação de contexto extraída dos segmentos circundantes e sobre os oradores.

Tendo em conta as características do corpus DIHANA e as melhores abordagens para o reconhecimento automático de actos de diálogo independentes do domínio e com apenas uma etiqueta, neste artigo exploramos diferentes aspectos relacionados com a tarefa. Em primeiro lugar, avaliamos se essas abordagens têm um desempenho semelhante numa língua diferente do inglês, utilizando-as para prever as etiquetas independentes do domínio do nível superior. Em seguida, exploramos a sua aplicabilidade nos cenários de classificação multi-etiqueta colocados pelos restantes níveis. Para além disso, uma vez que esses níveis se referem a diferentes aspectos específicos da tarefa, também avaliamos como a informação de contexto extraída dos segmentos anteriores influencia a capacidade de prever cada um desses aspectos. Da mesma forma, avaliamos como essa habilidade é influenciada por informação relativa aos níveis superiores da hierarquia. Por fim, exploramos a combinação hierárquica das melhores abordagens para cada nível e comparamos o seu desempenho com o da abordagem plana que foi utilizada em estudos anteriores sobre o mesmo corpus.

No resto do artigo, começamos por fornecer uma visão geral sobre o trabalho relacionado na Secção 2. Nesse sentido, começamos por fornecer uma visão geral sobre corpora para o reconhecimento de actos de diálogo na Secção 2.1. Em seguida, discutimos as melhores abordagens para o reconhecimento de actos de diálogo na Secção 2.2. Adicionalmente, resumimos estudos anteriores sobre o reconhecimento de actos de diálogo em dados em espanhol na Secção 2.3. Após essa discussão, na Secção 3, descrevemos a nossa configuração experimental. Começamos por descrever o corpus DIHANA e as suas anotações de actos de diálogo na Secção 3.1. A Secção 3.2 apresenta a arquitectura genérica das redes utilizadas nas

nossas experiências e descreve o que muda entre cada uma delas. Por fim, a Secção 3.3 descreve os procedimentos de treino e avaliação de acordo com o nível da hierarquia em foco. Os resultados alcançados pelas nossas experiências em cada um desses níveis, assim como na sua combinação, são apresentados e discutidos na Secção 4. Por fim, a Secção 5 apresenta as conclusões mais importantes que podem ser tiradas das experiências descritas neste artigo e fornece indicadores para trabalho futuro.

2 Trabalho Relacionado

Como mencionado anteriormente, o reconhecimento automático de actos de diálogo é uma tarefa que tem sido amplamente explorada ao longo dos anos em múltiplos corpora com diferentes características e usando uma grande variedade de técnicas de aprendizagem clássicas, desde Modelos de Markov Ocultos (HMMs) (Stolcke et al., 2000) até Máquinas de Vectores de Suporte (SVMs) (Gambäck et al., 2011). O artigo de Král & Cerisara (2010) fornece uma visão geral sobre a maioria dessas abordagens. No entanto, recentemente, a maioria das abordagens baseia-se em diferentes arquitecturas de DNNs. Abaixo, apresentamos um sumário dessas abordagens. Para além disso, uma vez que o nosso estudo se foca no corpus DIHANA (Benedí et al., 2006), temos também uma subsecção dedicada a abordagens aplicadas no reconhecimento de actos de diálogo em dados em espanhol. No entanto, antes de discutirmos abordagens, fornecemos uma visão geral sobre corpora existentes para o reconhecimento de actos de diálogo.

2.1 Corpora para Reconhecimento de Actos de Diálogo

Vários corpora foram anotados com actos de diálogo. A Tabela 1 apresenta um conjunto não exaustivo desses corpora e das suas características. Podemos ver que múltiplos domínios, linguagens e tipos de interação são cobertos, o que permite a avaliação das capacidades de generalização das abordagens de reconhecimento de actos de diálogo para múltiplos cenários. No entanto, por outro lado, os conjuntos de etiquetas utilizados não são padronizados entre corpora. De facto, existem até conjuntos distintos de etiquetas para o mesmo corpus. Isso significa que esses conjuntos foram desenvolvidos com diferentes objectivos e têm diferentes hierarquias e níveis de abstracção, o que dificulta a realização de experiências de generalização entre corpora. Isto é particularmente problemático quando os conjuntos de etiquetas usados são dependentes do domínio, uma vez que não podem ser aplicados a corpora noutros domínios.

Em relação a conjuntos alternativos de etiquetas para o mesmo corpus, enquanto os dos corpora SwDA, ICSI Meeting Recorder Dialog Act Corpus (MRDA) e CallHome Spanish (CHS) são apenas versões comprimidas dos conjuntos originais, os dois conjuntos de etiquetas usados para anotar o corpus VERBMOBIL são disjuntos. Para além disso, o primeiro inclui etiquetas dependentes do domínio (Jekat et al., 1995), enquanto o segundo é completamente independente do domínio (Alexandersson et al., 1998).

Múltiplos corpora têm conjuntos de etiquetas complementares que se referem a diferentes aspectos. Por exemplo, os corpora MRDA, DIHANA e NESPOLE têm um conjunto de etiquetas genéricas que podem ser especializadas usando etiquetas de outros conjuntos. No entanto, enquanto no primeiro caso as etiquetas especializadas ainda são independentes do domínio, nos dois restantes as etiquetas genéricas são complementadas com informação específica do domínio a diferentes níveis. No corpus DIME, os dois conjuntos de etiquetas referem-se a diferentes aspectos do diálogo, nomeadamente, definição de obrigações e estabelecimento de uma base comum. Por último, o corpus LEGO tem conjuntos de etiquetas independentes para os segmentos do utilizador e do sistema.

Numa tentativa de padronizar a anotação de actos de diálogo e, consequentemente, estabelecer uma base para estudos mais comparáveis na área, Bunt et al. (2012) definiram a norma ISO 24617-2. De acordo com esta norma, as anotações de actos de diálogo devem ser realizadas em segmentos funcionais, em vez de em turnos ou frases (Carroll & Tanenhaus, 1978). Para além disso, a anotação de cada segmento não consiste apenas numa etiqueta, mas sim numa estrutura complexa contendo informação sobre os participantes, relações com outros segmentos funcionais, a dimensão semântica do acto de diálogo, a sua função comunicativa e qualificadores opcionais sobre certeza, condicionalidade, parcialidade e sentimento. No entanto, anotar todos estes aspectos é um processo exaustivo e, consequentemente, a quantidade de dados anotados de acordo com a norma é ainda reduzida e, em muitos casos, nem todos os aspectos são considerados (Petukhova et al., 2014; Bunt et al., 2016; Ribeiro et al., 2016).

Como mencionado anteriormente, os estudos mais recentes sobre o reconhecimento automático

Corpus	Interacção	Domínio	Língua	Segmentos	Etiquetas	DD
SwDA (Jurafsky et al., 1997)	Humanos	Aberto	Inglês	220k	41 - 44	N
MRDA (Shriberg et al., 2004)	Humanos	Reuniões	Inglês	106k	5 / 11 + 39	N
AMI (Carletta et al., 2005)	Humanos	Reuniões	Inglês	102k	15	N
VERBMOBIL (Kay et al., 1992)	Humanos	Horários	Múltiplas	59k	42 / 33	M
CHS (Levin et al., 1998)	Humanos	Aberto	Espanhol	45k	10 / 37	N
DSTC4 (Kim et al., 2017)	Humanos	Viagens	Inglês	31k	89	S
MapTask (Anderson et al., 1991)	Humanos	Mapas	Inglês	27k	12	N
DIHANA (Benedí et al., 2006)	WoZ	Comboios	Espanhol	23k	11 + 10 + 13	M
LEGO (Schmitt et al., 2012)	Máquina	Autocarros	Inglês	14k	22 + 28	S
NESPOLE (Costantini et al., 2002)	Humanos	Viagens	Múltiplas	8k	67 + 91	M
DIME (Villaseñor et al., 2001)	WoZ	Cozinhas	Espanhol	5k	15 + 15	M

Tabela 1: Corpora anotados com actos de diálogo, ordenados por número aproximado de segmentos. A coluna referente à interacção diz se os diálogos são entre humanos ou existe um sistema de diálogo envolvido. No último caso, são distinguidos os cenários que usam o método WoZ daqueles que envolvem interacção real com uma máquina. Na coluna referente ao número de etiquetas, os símbolos / e - referem-se a conjuntos alternativos de etiquetas, enquanto o símbolo + se refere a diferentes níveis de anotação. A última coluna diz se o conjunto de etiquetas é dependente do domínio (S), independente do mesmo (N), ou se existem etiquetas de ambos os tipos (M).

de actos de diálogo utilizam diferentes arquitecturas de DNN. Tais abordagens requerem grandes quantidades de dados para serem treinadas. Consequentemente, a identificação automática de actos de diálogo como definidos pela norma ISO só foi abordada num conjunto reduzido de estudos (Ribeiro et al., 2015; Mezza et al., 2018). Por outro lado, o corpus SwDA é o mais explorado para a tarefa, uma vez que é aquele que possui o maior número de segmentos anotados, os seus diálogos cobrem múltiplos domínios e seu conjunto de etiquetas é independente do domínio. Por isso, é esperado que as conclusões tiradas de experiências sobre este corpus generalizem bem para outros cenários.

2.2 Estado da Arte em Reconhecimento de Actos de Diálogo

As abordagens com melhor desempenho na tarefa de reconhecimento de actos de diálogo são baseadas em DNNs. Por isso, nesta secção, focamo-nos em estudos que usam essa abordagem. Pelo que sabemos, o primeiro desses estudos foi o de Kalchbrenner & Blunsom (2013). O método descrito utiliza uma abordagem baseada em CNNs para gerar a representação de um segmento a partir da representação das suas palavras na forma de *embeddings* inicializados aleatoriamente. Em seguida, é usado um modelo de discurso baseado em RNNs que combina a sequência de representações de segmentos com informação sobre os oradores e produz a sequência de actos de diálogo correspondente. Ao limitar o modelo de discurso para considerar informação de apenas dois segmentos anteriores, esta abordagem alcançou uma taxa de acerto de 73,9 % no corpus SwDA.

Lee & Dernoncourt (2016) compararam o desempenho de uma unidade recorrente de Longa Memória de Curto Prazo (LSTM) com o de uma CNN para gerar representações de segmentos a partir da representação das suas palavras na forma de *embeddings* pré-treinados. Para identificar os actos de diálogo correspondentes, as representações de segmentos são passadas por uma rede totalmente ligada com duas camadas, na qual a primeira normaliza as representações e a segunda seleciona a classe com maior probabilidade. Nas experiências realizadas, a abordagem baseada em CNNs levou consistentemente a resultados semelhantes ou melhores do que aqueles da abordagem baseada em LSTM. A arquitectura foi também adaptada para fornecer informações de contexto a dois níveis e de até dois segmentos anteriores. O primeiro nível refere-se à concatenação das representações dos segmentos precedentes com a do segmento atual antes de o fornecer à rede totalmente ligada. O segundo refere-se à concatenação das representações normalizadas antes de serem fornecidas à camada de saída. Esta abordagem alcançou uma taxa de acerto de 65,8% no corpus Dialog State Tracking Challenge 4 (DSTC4), 84,6% no corpus MRDA com cinco classes (Ang et al., 2005) e 71,4% no corpus SwDA. No entanto, a influência da informação de contexto variou entre corpora.

Ji et al. (2016) exploraram a combinação de aspectos positivos de redes neuronais e modelos gráficos probabilísticos. Eles usaram um Modelo de Língua com Relações de Discurso (DRLM) que combina um Modelo de Língua Baseado em RNNs (RNNLM) (Mikolov et al., 2010), para modelar a sequência de palavras no diálogo, com um

modelo de variável latente sobre a estrutura do discurso, para modelar relações entre segmentos adjacentes que, neste contexto, representam os actos de diálogo. Desta forma, o modelo pode prever palavras usando representações vectoriais treinadas de forma discriminativa enquanto mantém uma representação probabilística de um elemento linguístico alvo, como o acto de diálogo. Para funcionar como um classificador de actos de diálogo, o modelo foi treinado para maximizar a probabilidade condicional de uma sequência de actos de diálogo dada uma sequência de segmentos, alcançando uma taxa de acerto de 77,0% no corpus SwDA.

Tran et al. (2017b) usaram uma RNN hierárquica com um mecanismo de atenção para prever as classificações de actos de diálogo de um diálogo inteiro. O modelo é hierárquico, uma vez que inclui uma RNN ao nível do segmento, para gerar a sua representação a partir das suas palavras, e outra para gerar a sequência de etiquetas de acto de diálogo a partir da sequência de representações de segmento. O mecanismo de atenção está entre os dois, uma vez que usa informações da RNN ao nível do diálogo para identificar as palavras mais importantes no segmento actual e filtrar a sua representação. Usando esta abordagem eles alcançaram uma taxa de acerto de 74,5% no corpus SwDA e 63,3% no corpus HCRC Map Task Corpus (MapTask). Mais tarde, o desempenho no corpus SwDA foi melhorado para 75,6% usando um método baseado na propagação de informação de incerteza sobre as previsões anteriores (Tran et al., 2017c). Para além disso, utilizando mecanismos de atenção aplicados às células das camadas recorrentes no contexto de um modelo generativo, alcançaram uma taxa de acerto de 74,2% no corpus SwDA e 65,94% no corpus MapTask (Tran et al., 2017a).

Os estudos referidos anteriormente exploraram a utilização de uma única camada recorrente ou convolucional para gerar a representação do segmento a partir das suas palavras. No entanto, as abordagens com melhor desempenho na tarefa utilizam múltiplas dessas camadas. Por um lado, Khanpour et al. (2016) alcançaram os seus melhores resultados usando uma representação de segmento gerada pela concatenação das saídas de uma pilha de 10 unidades LSTM. Deste modo, o modelo é capaz de capturar relações de longa distância entre palavras. Por outro lado, Liu et al. (2017) geraram a representação do segmento combinando as saídas de três CNNs paralelas com diferentes tamanhos de janela de contexto, para capturar diferentes padrões funcionais. Em ambos os casos, representações das palavras na forma de *embeddings* pré-treinados foram usadas como entrada para a rede. Em geral, a partir dos resultados reportados, não é possível afirmar qual é a abordagem de representação de segmento com melhor desempenho, uma vez que a avaliação foi realizada em diferentes subconjuntos do corpus SwDA. Ainda assim, Khanpour et al. (2016) atingiram uma taxa de acerto de 73,9% no conjunto de validação e 80,1% no conjunto de teste, enquanto Liu et al. (2017) atingiram taxas de acerto de 74,5% e 76,9% nos dois conjuntos utilizados para avaliar as suas experiências. Para além disso, Khanpour et al. (2016) atingiram uma taxa de acerto de 86,8% no corpus MRDA.

Liu et al. (2017) exploraram também o uso de informação de contexto sobre mudança de orador e extraída dos segmentos circundantes. Para fornecer informação sobre a mudança de orador limitaram-se a adicionar um valor binário à representação do segmento, que indica se o orador mudou em relação ao segmento anterior. Já em relação a informação dos segmentos circundantes, exploraram o uso de modelos de discurso, assim como de abordagens que concatenam a informação de contexto directamente na representação do segmento. Os modelos de discurso tornam o modelo hierárquico, gerando uma sequência de classificações de actos de diálogo a partir da sequência de representações de segmento. Assim, ao prever a classificação de um segmento, aqueles que o circundam também são levados em conta. No entanto, quando o modelo de discurso é baseado numa CNN ou numa unidade LSTM bidireccional, ele considera informação de segmentos futuros, que não estão disponíveis para um sistema de diálogo. Ainda assim, mesmo tendo em conta informação futura, as abordagens baseadas em modelos de discurso tiveram pior desempenho do que aquelas que concatenaram a informação de contexto directamente na representação do segmento. Nesse aspecto, fornecer essa informação na forma das classificações de acto de diálogo dos segmentos circundantes levou a melhores resultados do que utilizar as suas palavras, mesmo quando essas classificações foram obtidas automaticamente. Esta conclusão está alinhada com o que tínhamos mostrado no nosso estudo anterior, utilizando SVMs (Ribeiro et al., 2015). Para além disso, ambos os estudos demonstraram que, como esperado, o primeiro segmento anterior é o mais importante e que a influência diminui com a distância. Usando as etiquetas de referência de três segmentos anteriores, os resultados nos dois conjuntos usados para avaliar a abordagem melhoraram para 79,6% e 81,8%, respectivamente.

É importante fazer algumas observações sobre tokenização e representação de *tokens*. Em todos os estudos descritos anteriormente, a tokenização foi realizada no nível da palavra. Para além disso, com excepção do primeiro estudo (Kalchbrenner & Blunsom, 2013), em que foram utilizadas representações na forma de *embeddings* inicializados aleatoriamente, e dos realizados por Tran et al. (2017a,b,c), para os quais a abordagem de representação não é revelada nos artigos, a representação dessas palavras é feita na forma de *embeddings* pré-treinados. Khanpour et al. (2016) compararam a performance utilizando *embeddings* treinados usando os métodos Word2Vec (Mikolov et al., 2013) e Vectores Globais para Representação de Palavras (GloVe) (Pennington et al., 2014) em múltiplos corpora. Embora ambas as abordagens capturem informação relativa a palavras que aparecem juntas frequentemente, os melhores resultados foram alcançados usando a abordagem Word2Vec. Em termos de dimensionalidade, Khanpour et al. (2016) alcançou os melhores resultados ao usar *embeddings* com 150 dimensões. No entanto, outros estudos (Lee & Dernoncourt, 2016; Liu et al., 2017) usam *embeddings* com 200 dimensões, sendo que este não foi um dos valores comparados para a dimensionalidade.

As abordagens descritas em todos os estudos referidos anteriormente realizam tokenização ao nível da palavra. No entanto, num estudo recente (Ribeiro et al., 2018), mostrámos que também existem pistas importantes para a detecção de intenção a um nível sub-palavra, que só podem ser capturadas quando se usa uma tokenização mais granular, como, por exemplo, ao nível do carácter (Ribeiro et al., 2018). As pistas a esse nível referem-se principalmente a aspectos relativos à morfologia das palavras, tais como lemas e afixos. Para capturar essa informação, nós adaptámos a abordagem de representação de segmento baseada em CNNs descrita por Liu et al. (2017) para usar caracteres em vez de palavras. Dessa forma, pudemos explorar janelas de contexto de diferentes tamanhos para capturar diferentes aspectos morfológicos. Neste sentido, os nossos melhores resultados foram alcançados quando utilizámos três CNNs paralelas com janelas de tamanho três, cinco e sete, que são capazes de capturar afixos, lemas e relações entre palavras, respectivamente. Usando essa abordagem, obtivemos taxas de acerto de 76,88% e 73,22% nos conjuntos de validação e teste do corpus SwDA, respectivamente. Estes resultados são semelhantes aos da abordagem ao nível da palavra. No entanto, a combinação dos dois níveis melhorou os resultados para 78,0% e 74,0%, respectivamente, o que mostra que estes capturam informação complementar. Por fim, ao incluir informação de contexto de três segmentos anteriores, melhorámos os resultados para 82,0% no conjunto de validação e 79,0% no conjunto de teste.

2.3 Reconhecimento de Actos de Diálogo em Dados em Espanhol

A investigação sobre o reconhecimento de actos de diálogo em dados em espanhol tem sido realizada principalmente em dois corpora — DIHANA e CHS. Em ambos, os diálogos são telefónicos e espontâneos. No entanto, tal como mostrado na Tabela 1, enquanto os diálogos do primeiro são entre humanos e um sistema de diálogo, os do segundo são entre humanos. Para além disso, enquanto o corpus CHS está anotado com etiquetas independentes do domínio e da tarefa, o corpus DIHANA está anotado segundo um esquema hierárquico com três níveis, em que o primeiro se refere ao acto de diálogo genérico e independente do domínio e os restantes o complementam com informação específica da tarefa. Existe também uma série de trabalhos sobre reconhecimento de actos de diálogo no corpus DIME (Coria & Pineda, 2005, 2006, 2009). No entanto, estes focam-se em usar informação prosódica para prever subconjuntos específicos dos actos de diálogo relacionados com obrigações e estabelecimento de uma base comum com que o corpus está anotado. Uma vez que o nosso trabalho se foca no reconhecimento de actos de diálogo a partir de informação textual, apenas vamos descrever mais detalhadamente os estudos sobre os dois primeiros corpora.

Os primeiros estudos em reconhecimento de actos de diálogo no corpus DIHANA usaram HMMs, quer baseados em informação prosódica (Tamarit & Martínez-Hinarejos, 2008) — energia e frequência fundamental —, quer textual (Martínez-Hinarejos et al., 2008) — n-gramas de palavras. O primeiro atingiu uma taxa de acerto de 60,70% no primeiro nível, enquanto o segundo alcançou 93,40% na combinação dos dois primeiros níveis e 89,70% na combinação de todos os níveis. Este segundo estudo e um outro mais recente (Martínez-Hinarejos et al., 2015) também exploraram o reconhecimento de actos de diálogo em diálogos não segmentados à priori, usando transdutores de n-gramas. No entanto, nesses casos, o foco foi no processo de segmentação e as abordagens de classificação actos de diálogo não diferiram das anteriores. Por fim, os melhores resultados nos diálogos segmentados manualmente foram obtidos usando uma abordagem baseada

em SVMs aplicados a n-gramas de palavras, informação sobre a presença de palavras de pergunta e pontuação, e informação de contexto de três segmentos anteriores (Gambäck et al., 2011). Para além disso, foi também aplicada uma abordagem de aprendizagem activa para reduzir a quantidade de dados necessários para o treino, alcançando uma taxa de acerto de 94,08% na combinação dos dois primeiros níveis e 90,97% na combinação de todos os níveis.

Tal como no corpus DIHANA, os primeiros estudos em reconhecimento de actos de diálogo no corpus CHS também usaram HMMs com diferentes tipos de n-grama (Levin et al., 1999; Ries, 1999). O segundo estudo referido melhorou os resultados ao combinar os HMMs com redes neuronais aplicadas a unigramas etiquetas de Parte do Discurso (POS), alcançando uma taxa de acerto de 76.1%. A tarefa também foi abordada usando Análise de Semântica Latente (LSA) em três estudos diferentes (Serafin et al., 2003; Serafin & Di Eugenio, 2004; Di Eugenio et al., 2010). O primeiro usou não só LSA básica, como também múltiplas adaptações baseadas em aglomeração e na incorporação de informação relativa aos actos de diálogo anteriores. No entanto, não foi observada uma melhoria significativa em relação à LSA básica, que alcançou uma taxa de acerto de 65,36% no conjunto com 37 etiquetas 68,91% na sua versão comprimida, com 10 etiquetas. Por outro lado, os restantes estudos exploraram o uso de informação sobre múltiplas características sintáticas e relacionadas com o diálogo, atingindo resultados superiores aos obtidos usando LSA básica, até um máximo de 77,74% e 81,27%, respectivamente. No último estudo, esses resultados foram ainda melhorados para 80,34% e 82,88%, através da aplicação de uma abordagem de aprendizagem baseada em instâncias, mais especificamente, k-Vizinhos Mais Próximos (k-NN), aos espaços semânticos obtidos através da LSA. No entanto, em ambos os casos, as melhorias foram alcançadas utilizando informação relativa ao objectivo do diálogo, ou seja, a intenção genérica por trás de todo o diálogo, e sobre se o orador está a tomar a iniciativa, ou apenas a responder ou a acompanhar o outro orador. Embora em geral o objectivo do diálogo seja conhecido, existem alguns casos em que um sistema de diálogo não tem essa informação. Para além disso, identificar se um orador está a tomar a iniciativa, a responder ou a acompanhar o outro orador pode ser visto como uma simplificação da tarefa de reconhecimento de actos de diálogo. Logo, não é justo usar essa informação caso esta não seja obtida automaticamente também. Por fim, o corpus CHS também foi explorado em experiências de adaptação a diferentes domínios para classificação de actos de diálogo usando um conjunto reduzido de classes (Margolis et al., 2010).

3 Configuração Experimental

Queremos avaliar se as abordagens com melhor desempenho descritas na secção anterior têm um desempenho semelhante numa língua diferente do inglês. Para além disso, queremos explorar a sua aplicabilidade nos cenários de classificação multi-etiqueta colocados pelos dois níveis inferiores das anotações de actos do diálogo do corpus DIHANA. Como esses níveis se referem a diferentes aspectos específicos da tarefa, também avaliamos como a informação de contexto extraída dos segmentos anteriores influencia a capacidade de prever cada um desses aspectos. Da mesma forma, avaliamos como essa capacidade é influenciada por informação dos níveis acima na hierarquia. Por fim, queremos avaliar se a combinação hierárquica das melhores abordagens para cada nível é capaz de superar a abordagem plana utilizada em estudos anteriores sobre o mesmo corpus.

Nesta secção descrevemos a nossa configuração experimental, começando com uma descrição do corpus DIHANA e das suas anotações de acto de diálogo. Em seguida, apresentamos a arquitectura genérica usada nas nossas experiências e explicamos como ela muda de acordo com o aspecto e as características do nível em foco, especialmente considerando as diferenças entre os cenários de classificação com etiqueta única e os de classificação multi-etiqueta. Por fim, descrevemos as nossas abordagens de treino e avaliação, incluindo as diferenças nas métricas usadas para problemas com etiqueta única e multi-etiqueta.

3.1 Corpus

O corpus DIHANA (Benedí et al., 2006) consiste em 900 diálogos telefónicos entre 225 humanos e um sistema de diálogo que fornece informação sobre comboios, simulado usando o método WoZ. Existem 6.280 turnos de utilizadores e 9.133 turnos do sistema, com um vocabulário de 823 palavras e um total de 48.243 palavras. Os turnos foram transcritos, segmentados e anotados manualmente com actos de diálogo (Alcácer et al., 2005). O número total de segmentos anotados é 23.547, 9.715 dos quais são de utilizadores e 13.832 do sistema. A Figura 1 mostra um exemplo de um diálogo anotado.

SISTEMA: Bienvenido al servicio de informacion de trenes ¿En qué puedo ayudarle?
(Bem-vindo ao serviço de informação de comboios. Como posso ajudá-lo?)
Nível 1: *Apertura*, Nível 2: *Nula*, Nível 3: *Nula*
UTILIZADOR: Quería ir a La Coruña *(Queria ir à Corunha)*
Nível 1: *Pregunta*, Nível 2: *Hora Salida*, Nível 3: *Destino*
UTILIZADOR: y quería saber los horarios que hay de los trenes para el lunes.
(e gostaria de saber o horário dos comboios à segunda-feira.)
Nível 1: *Pregunta*, Nível 2: *Hora Salida*, Nível 3: *Día*
SISTEMA: ¿Quiere viajar desde desde La Coruña?
(Quer viajar a partir da Corunha?)
Nível 1: *Confirmación*, Nível 2: *Origen*, Nível 3: *Origen*
UTILIZADOR: No. *(Não.)*
Nível 1: *Negación*, Nível 2: *Origen*, Nível 3: *Nula*
UTILIZADOR: Desde Bilbao. *(A partir de Bilbau.)*
Nível 1: *Respuesta*, Nível 2: *Hora Salida*, Nível 3: *Origen*
SISTEMA: Quiere horarios de trenes a La Coruña
(Quer horários de comboios para a Corunha)
Nível 1: *Confirmación*, Nível 2: *Hora Salida*, Nível 3: *Destino*
SISTEMA: ¿desde Bilbao? *(a partir de Bilbau?)*
Nível 1: *Confirmación*, Nível 2: *Origen*, Nível 3: *Origen*
UTILIZADOR: Sí. *(Sim.)*
Nível 1: *Afirmación*, Nível 2: *Origen*, Nível 3: *Nula*
SISTEMA: Le consulto horarios de trenes con salida en Bilbao destino La Coruña saliendo el lunes, día 16 de febrero de 2004.
(Estou à procura de comboios de Bilbau para a Corunha na segunda-feira, 16 de Fevereiro de 2004.)
Nível 1: *Confirmación*, Nível 2: *Hora Salida*, Nível 3: *Destino, Día, Origen*
SISTEMA: Un momento por favor. *(Um momento por favor.)*
Nível 1: *Espera*, Nível 2: *Nula*, Nível 3: *Nula*
SISTEMA: El único tren que realiza el trayecto es un Diurno que sale a las 9 y 25 de la mañana.
(O único comboio que faz esse percurso é um Diurno que parte às 9 e 25 da manhã.)
Nível 1: *Respuesta*, Nível 2: *Hora Salida*, Nível 3: *Hora Salida,Número Trenes,Tipo Tren*
SISTEMA: ¿Desea algo más? *(Precisa de mais alguma coisa?)*
Nível 1: *Nueva Consulta*, Nível 2: *Nula*, Nível 3: *Nula*
UTILIZADOR: Sí, ¿me puede decir el precio del billete?
(Sim. Pode dizer-me o preço do bilhete?)
Nível 1: *Pregunta*, Nível 2: *Precio*, Nível 3: *Nula*
SISTEMA: El precio de ese tren en clase turista es 36 euros el billete de ida. Recuerde que el billete de ida y vuelta siempre supone un ahorro del 20 por ciento.
(O preço do bilhete de ida nesse comboio é de 36 euros em classe económica. Lembre-se que bilhetes de ida e volta têm um desconto de 20%.)
Nível 1: *Respuesta*, Nível 2: *Precio*, Nível 3: *Clase Billete, Precio*
SISTEMA: ¿Desea algo más? *(Precisa de mais alguma coisa?)*
Nível 1: *Nueva Consulta*, Nível 2: *Nula*, Nível 3: *Nula*
UTILIZADOR: No. Creo que no. *(Não. Penso que não.)*
Nível 1: *Encerramento*, Nível 2: *Nula*, Nível 3: *Nula*
SISTEMA: Gracias por utilizar este servicio. Feliz viaje.
(Obrigado por usar este serviço. Boa viagem.)
Nível 1: *Encerramento*, Nível 2: *Nula*, Nível 3: *Nula*

Figura 1: Um diálogo do corpus DIHANA. Cada segmento é seguido pela sua tradução em português, entre parêntesis, e pelas suas etiquetas de actos de diálogo em cada um dos três níveis.

As anotações de actos de diálogo são decompostas hierarquicamente em três níveis (Martínez-Hinarejos et al., 2002). Enquanto o primeiro nível representa a intenção genérica do segmento, independente de detalhes relativos ao domínio e à tarefa, os restantes representam informação específica da tarefa. O primeiro nível tem 11 etiquetas, distribuídas de acordo com a Tabela 2. Nessa tabela podemos ver que duas das etiquetas são exclusivas a segmentos de utilizadores — *Aceitação* e *Rejeição* — e quatro a segmentos do sistema — *Abertura*, *Espera*, *Nova Consulta* e *Confirmação*. Para além disso, a etiqueta mais comum, *Pergunta*, cobre 27% dos segmentos.

Etiqueta	**U**	**S**	**T**	**%**
Pregunta *(Pergunta)*	5.474	864	6.338	27
Respuesta *(Resposta)*	1.839	2.446	4.285	18
Confirmación *(Confirmação)*	0	3.629	3.629	15
Nueva Consulta *(Nova Consulta)*	0	2.474	2.474	11
Espera *(Espera)*	0	1.948	1.948	8
Cierre *(Encerramento)*	927	900	1.827	8
Afirmación *(Aceitação)*	990	0	990	4
Apertura *(Abertura)*	0	900	900	4
No Entendido *(Não Percebido)*	4	653	657	3
Negación *(Rejeição)*	340	0	340	1
Indefinida *(Indefinida)*	141	18	159	1

Tabela 2: Distribuição das etiquetas de Nível 1 no corpus. A tradução em português de cada etiqueta está entre parêntesis. As colunas identificadas como U, S e T referem-se ao número de segmentos de utilizador, sistema e total anotados com a etiqueta.

Embora partilhem a maioria das etiquetas, os dois níveis inferiores da hierarquia focam-se em diferentes tipos de informação específica da tarefa. Enquanto o segundo nível está relacionado com o tipo de informação que é implicitamente focado pelo segmento, o terceiro nível está relacionado com o tipo de informação que é explicitamente referido no segmento. A título ilustrativo, vamos olhar para o segmento *"Estou à procura de comboios de Bilbau para a Corunha na segunda-feira, 16 de Fevereiro de 2004."*, extraído do diálogo da Figura 1. Uma vez que o segmento revela a intenção de encontrar um horário de comboio, este tem *Hora de Partida* como etiqueta de Nível 2. No entanto, como esse horário de partida não é explicitamente referido no segmento, essa etiqueta não faz parte das suas etiquetas de Nível 3. Por outro lado, o segmento refere explicitamente um local de partida, um destino e uma data. Logo, tem as etiquetas de Nível 3 correspondentes — *Origem*, *Destino* e *Dia*.

A distribuição das etiquetas de ambos os níveis é mostrada na Tabela 3. Podemos ver que existem 10 etiquetas comuns e três adicionais no Nível 3 — *Número de Ordem*, *Número de Comboios* e *Tipo de Viagem*. Para além disso, ambos os níveis têm a etiqueta *Nula*, que representa a ausência de etiqueta nesse nível. Neste sentido, podemos ver que apenas 63% dos segmentos têm etiquetas de Nível 2, e que a percentagem é ainda menor, 52%, quando se consideram etiquetas de Nível 3. Isto deve-se principalmente ao facto de que segmentos etiquetados como *Abertura*, *Encerramento*, *Indefinida*, *Não Entendido*, *Espera*, e *Nova Consulta* no primeiro nível não podem ter etiquetas nos restantes níveis. Por fim, é importante referir que cada segmento tem uma e apenas uma etiqueta de Nível 1, mas pode ter várias etiquetas de Nível 2 e Nível 3.

Como observação final, é importante referir que algumas etiquetas de Nível 2 — *Duração*, *Classe* e *Serviço* — e de Nível 3 — *Serviço* e *Duração* — ocorrem apenas em 0,1% dos segmentos ou menos. Por isso, essas etiquetas são especialmente difíceis de prever usando métodos de aprendizagem automática que se focam em maximizar a taxa de acerto no corpus como um todo.

3.2 Arquitectura da Rede

Uma vez que queremos avaliar o desempenho de diferentes abordagens baseadas em DNNs no reconhecimento de actos de diálogo no DIHANA, precisamos de definir uma base comum para comparação. Para tal, usamos uma arquitectura de rede genérica, mostrada na Figura 2, baseada naquelas das abordagens com melhor desempenho referidas na Secção 2.2. Usando esta arquitectura, a abordagem para identificar os actos de diálogo de um segmento é a seguinte: Primeiro, o segmento é dividido em *tokens*, que são passados por uma camada de *embedding* para gerar as suas representações nessa forma. Em seguida, a sequência de *embeddings* é passada para a abordagem de representação do segmento. A representação obtida pode ser complementada com informação de contexto, antes de ser passada por uma camada de redução de dimensionalidade. Por fim, a representação reduzida é passada para a camada de saída, que gera a classificação de actos de diálogo do segmento. A motivação para cada um destes passos e a forma como as suas características variam de acordo com o nível da hierarquia em foco são descritas abaixo.

Nível 2					Nível 3				
Etiqueta	**U**	**S**	**T**	**%**	**Etiqueta**	**U**	**S**	**T**	**%**
Nulo *(Nula)*	1.923	6.893	8.816	37	Nulo *(Nula)*	2.954	8.317	11.271	48
Hora Salida *(Hora de Partida)*	3.309	3.523	7.432	32	Destino *(Destino)*	1.631	2.079	3.710	16
Precio *(Preço)*	2.071	1.267	3.338	14	Día *(Dia)*	1.881	1.778	3.659	16
Día *(Dia)*	1.026	923	1.949	8	Origen *(Origem)*	896	2.085	2.981	13
Origen *(Origem)*	477	480	957	4	Hora Salida *(Hora de Partida)*	692	1.633	2.325	10
Destino *(Destino)*	452	400	852	4	Número Trenes *(Número de Comboios)*	0	1.863	1.863	8
Tipo Tren *(Tipo de Comboio)*	317	226	543	2	Tipo Tren *(Tipo de Comboio)*	544	1.253	1.797	8
Hora Llegada *(Hora de Chegada)*	90	88	178	1	Número Orden *(Número de Ordem)*	84	950	1.034	4
Tiempo Recorrido *(Duração)*	14	15	29	0,1	Clase Billete *(Classe)*	129	766	895	4
Clase Billete *(Classe)*	15	12	27	0,1	Precio *(Preço)*	47	731	778	3
Servicio *(Serviço)*	3	5	8	0	Hora Llegada *(Hora de Chegada)*	199	490	689	3
					Tipo Viaje *(Tipo de Viagem)*	643	0	643	3
					Servicio *(Serviço)*	15	4	19	0,1
					Tiempo Recorrido *(Duração)*	0	14	14	0,1

Tabela 3: Distribuição das etiquetas de Nível 2 e Nível 3 no corpus. A tradução em português de cada etiqueta está entre parêntesis. As colunas identificadas como U, S e T referem-se ao número de segmentos de utilizador, sistema e total anotados com a etiqueta.

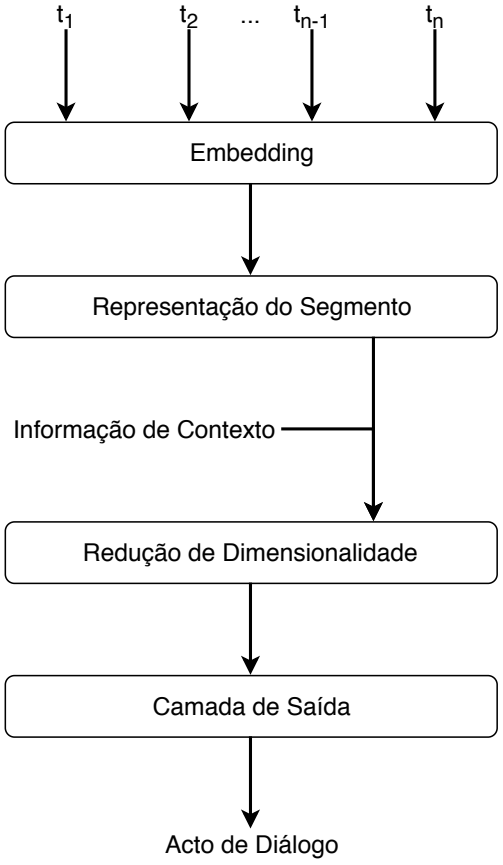

Figura 2: A arquitectura genérica das redes usadas nas nossas experiências. t_i corresponde ao i-ésimo *token*.

3.2.1 Embedding

A entrada da nossa rede é a sequência de *tokens* do segmento. De forma semelhante à maioria dos estudos anteriores sobre o reconhecimento de actos de diálogo, nós utilizamos tokenização no nível da palavra. Como mostrado no nosso estudo anterior (Ribeiro et al., 2018), o nível do carácter também é capaz de fornecer informação importante. No entanto, como forma de simplificação, não o incluímos neste estudo. Para além disso, ignoramos a pontuação, pois esta pode não estar disponível para um sistema de diálogo. Os tokens são então passados para a camada de *embedding* para serem transformados numa representação vectorial correspondente à sua posição no espaço de *embeddings*. Nas nossas experiências usamos *embeddings* pré-treinados usando o método Word2Vec (Mikolov et al., 2013) no Spanish Billion Words Corpus (Cardellino, 2016). Embora tenhamos explorado espaços de *embeddings* com diferentes dimensionalidades, apenas reportamos os resultados obtidos utilizando a dimensionalidade 200, tal como no estudo de Liu et al. (2017), uma vez que esta levou consistentemente a melhores resultados do que as dimensionalidades exploradas por Khanpour et al. (2016).

3.2.2 Representação do Segmento

Este passo gera uma representação vectorial do segmento através da combinação das representações dos seus *tokens*. Tal como referido na Secção 2.2, as abordagens com melhor desempenho no reconhecimento de actos de diálogo em dados em inglês diferem neste passo. Enquanto a abordagem de Khanpour et al. (2016) é baseada em RNNs, a de Liu et al. (2017) é baseada em CNNs. Ambas têm as suas vantagens, uma vez que enquanto a primeira se foca em capturar informação de sequências de *tokens* relevantes, a segunda foca-se no contexto que circunda cada *token* e, por isso, captura padrões relevantes.

Uma vez que os diferentes níveis da hierarquia de anotação de actos de diálogo do corpus DIHANA têm diferentes características, nós usamos ambas as abordagens nas nossas experiências para avaliar se existe uma com melhor desempenho em qualquer situação ou se existe uma dependência do nível em foco.

Tal como descrito na Secção 2.2, a abordagem baseada em RNNs de Khanpour et al. (2016) usa uma pilha de 10 unidades LSTM. A representação do segmento é dada pela concatenação das saídas das 10 unidades após processarem todos os *tokens* do segmento. Usar as saídas após o processamento de todos os *tokens* faz sentido, uma vez que estes são processados sequencialmente pelas unidades recorrentes e, por isso, essas saídas contêm informação de todo o segmento. Os resultados reportados neste artigo foram obtidos usando uma pilha de cinco Unidades Recorrentes *Gated* (GRUs) em vez de 10 LSTMs, uma vez que, nas nossas experiências preliminares, o desempenho foi semelhante, mas com um consumo de recursos significativamente menor. A Figura 3 mostra uma representação gráfica desta abordagem.

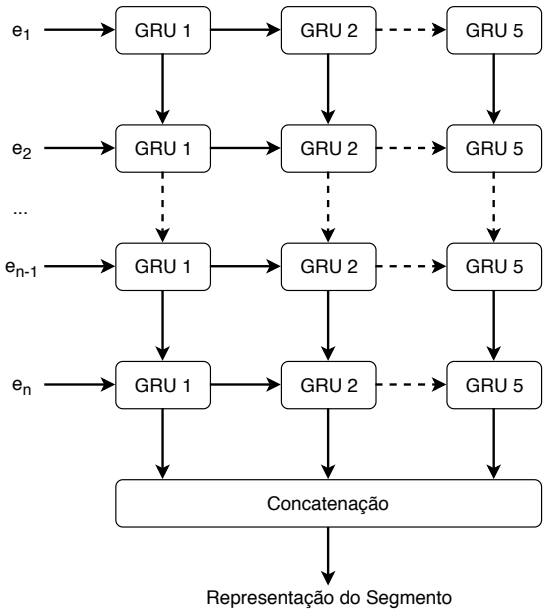

Figura 3: A abordagem de representação do segmento baseada em RNNs. e_i corresponde à representação do i-ésimo *token* na forma de *embedding*.

Também como descrito na Secção 2.2, a abordagem baseada em CNNs de Liu et al. (2017) usa três CNNs temporais paralelas com janelas de contexto com tamanho entre um e três, inclusive. Isto significa que a abordagem se foca em conjuntos de no máximo três palavras consecutivas. Um estudo anterior (Kim, 2014) usou janelas de contexto com tamanho entre três e cinco, de forma a capturar relações entre palavras mais distantes que eram relevantes para as tarefas exploradas. Tendo em conta a tarefa que estamos a explorar, as janelas de contexto mais relevantes dependem do nível em foco, uma vez que os actos de diálogo específicos da tarefa estão tipicamente relacionados com a presença de palavras específicas, enquanto os actos de diálogo genéricos estão mais relacionados com a estrutura do segmento e, consequentemente, com janelas mais largas. Para confirmar isto, usamos os dois conjuntos de janelas de contexto nas nossas experiências. As saídas das CNNs são filtradas usando uma operação de *max pooling* e são em seguida concatenadas para gerar a representação do segmento. A Figura 4 mostra uma representação gráfica desta abordagem.

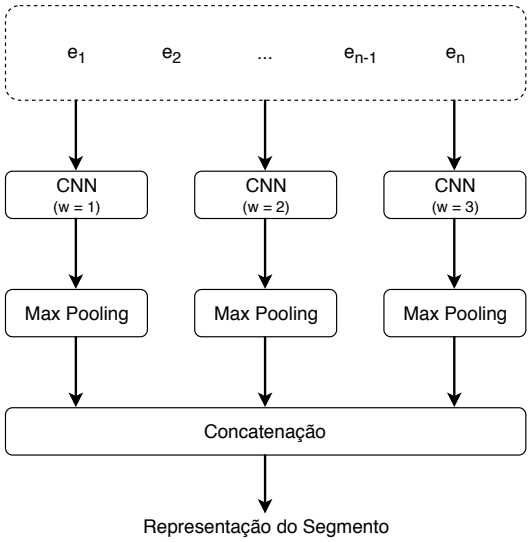

Figura 4: A abordagem de representação do segmento baseada em RNNs. e_i corresponde à representação do i-ésimo *token* na forma de *embedding*. O parâmetro w refere-se ao tamanho da janela de contexto.

3.2.3 Informação de Contexto

Vários estudos anteriores confirmaram a importância de informação de contexto extraída dos segmentos anteriores para a tarefa de reconhecimento de actos de diálogo (Ribeiro et al., 2015; Lee & Dernoncourt, 2016; Liu et al., 2017). Adicionalmente, esses estudos mostraram que a influência dos segmentos anteriores decresce com a distância e que as classificações de actos de diálogo desses segmentos são mais informativas que as suas palavras. Por isso, nas nossas

experiências fornecemos informação de contexto à rede usando a mesma abordagem baseada nas etiquetas dos segmentos anteriores usada no nosso estudo anterior no corpus SwDA (Ribeiro et al., 2015) e por Liu et al. (2017). Isto é, as etiquetas dos segmentos anteriores são transformadas numa representação vectorial única e concatenadas à representação do segmento. Tal como Liu et al. (2017), exploramos o uso de informação de contexto extraída de até um máximo de três segmentos anteriores, uma vez que o nosso estudo anterior mostrou que não existem melhorias significativas ao usar segmentos adicionais. No contexto de um sistema de diálogo que tenta identificar a intenção do seu interlocutor, este só tem acesso aos segmentos anteriores. Como tal, não usamos informação extraída de segmentos futuros nas nossas experiências. É importante referir que usamos as anotações manuais dos segmentos para fornecer a informação de contexto. Portanto, os resultados obtidos representam um tecto para o desempenho da abordagem. Optámos por não usar etiquetas obtidas automaticamente, uma vez que tanto o nosso estudo e o de Liu et al. (2017) mostraram que esta abordagem tem melhor desempenho que aquelas que usam as palavras de segmentos anteriores, mesmo quando as etiquetas são obtidas automaticamente. De acordo com esses estudos, é esperado que a taxa de acerto decresça cerca de dois pontos percentuais ao usar etiquetas obtidas automaticamente. No entanto, um sistema de diálogo está ciente dos actos de diálogo dos seus próprios segmentos. Como tal, só a classificação dos segmentos do utilizador está sujeita a erro, o que se espera que reduza o decréscimo da taxa de acerto. Ainda assim, como trabalho futuro, é importante avaliar qual o valor real desse decréscimo neste cenário.

Adicionalmente, uma vez que o corpus DIHANA está anotado com etiquetas de actos de diálogo hierárquicas, quando nos focamos num dado nível, exploramos também o uso de informação de contexto extraída dos níveis superiores, tanto relativamente ao segmento actual como aos anteriores. Para fornecer essa informação, usamos a mesma abordagem baseada em etiquetas usada para fornecer informação de contexto dos segmentos anteriores.

3.2.4 Redução de Dimensionalidade

Para evitar possíveis diferenças de resultados causadas pelo uso de representações de segmentos com diferente dimensionalidade, a nossa arquitectura inclui uma camada de redução de dimensionalidade que mapeia a representação do segmento, incluindo informação de contexto, num espaço com 100 dimensões. Deste modo, as diferenças de desempenho que possam ser observadas devem-se à natureza da abordagem de representação do segmento e à informação que esta é capaz de capturar e não a factores relacionados com a dimensionalidade. Para além disso, para reduzir a probabilidade de haver um sobreajustamento aos dados de treino, esta camada aplica também uma técnica de *dropout*, desactivando 50% dos neurónios durante a fase de treino.

3.2.5 Camada de Saída

A camada de saída mapeia a representação reduzida do segmento nas etiquetas de actos de diálogo correspondentes. Este processo é feito usando uma camada totalmente ligada com um número de neurónios igual ao número de etiquetas. Como cada segmento tem apenas uma etiqueta de Nível 1, usamos *softmax* como função de activação e a entropia cruzada categórica como função de custo. No entanto, essa abordagem não é válida para os restantes níveis, uma vez que estes permitem que um segmento tenha múltiplas etiquetas. Por isso, nesses casos, usamos a função de activação sigmoide e a entropia cruzada binária como função de custo que, tendo em conta a possibilidade de múltiplas etiquetas, é na verdade a função de custo de Hamming, apropriada para este tipo de problema (Díez et al., 2015). Em ambos os casos, por questões de desempenho, usamos o optimizador Adam (Kingma & Ba, 2015).

3.3 Treino e Avaliação

Para implementar as nossas redes usámos a API de alto nível Keras (Chollet et al., 2015) fornecida com a biblioteca TensorFlow (Abadi et al., 2016). Usámos uma metodologia de treino em lotes de tamanho 512 e interrompemos o treino após 10 épocas sem melhorias no conjunto de validação. Uma vez que existe algum não-determinismo envolvido, especialmente devido à execução em Unidade de Processamento Gráfico (GPU), os resultados apresentados na próxima secção referem-se à média (m) e ao desvio padrão (s) dos resultados obtidos em 10 execuções.

Para avaliar as nossas abordagens, fizemos uma validação cruzada com cinco partições, usando as partições definidas nos primeiros estudos sobre o corpus DIHANA (Tamarit & Martínez-Hinarejos, 2008; Martínez-Hinarejos et al., 2008). As métricas de avaliação utilizadas variam de acordo com o nível da hierarquia em foco. Uma vez que cada segmento tem apenas

uma etiqueta de Nível 1, neste caso lidamos com um problema de classificação de etiqueta única. Como tal, de forma semelhante a estudos anteriores sobre reconhecimento automático de actos de diálogo, avaliamos o desempenho usando a taxa de acerto (Acc). No entanto, essa não é a métrica mais adequada para os Níveis 2 e 3 da hierarquia, uma vez que estes colocam problemas de classificação multi-etiqueta. Por isso, avaliamos o desempenho nesses níveis usando as métricas adaptadas a cenários multi-etiqueta descritas por Sorower (2010). A métrica equivalente à taxa de acerto em cenários multi-etiqueta é a taxa de correspondência exacta (MR), definida como

$$\mathrm{MR} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Z_i), \tag{1}$$

onde Y_i é o conjunto de etiquetas de referência do exemplo i, Z_i é o conjunto de etiquetas previstas pelo classificador para o mesmo exemplo e I é a função indicadora. O problema desta métrica é que não considera acertos parciais, que são comuns em problemas de classificação multi-etiqueta. De forma a considerar esses casos, as métricas tradicionais para problemas de etiqueta única — taxa de acerto (Acc), precisão (P), sensibilidade (R) e medida-F (F_1) — são adaptadas da seguinte forma:

$$\mathrm{Acc} = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}, \tag{2}$$

$$P = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Z_i|}, \tag{3}$$

$$R = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Y_i|}, \tag{4}$$

$$F_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|}. \tag{5}$$

onde o operador $|X|$ é usado para obter a cardinalidade do conjunto X. Para além disso, tal como referido na Secção 3.2.5, a função de custo de Hamming (HL), que diz quantas vezes, em média, a relevância de um exemplo para uma etiqueta é incorrectamente prevista e é definida como

$$\mathrm{HL} = \frac{1}{n|L|}\sum_{i=1}^{n}\sum_{l\in L}[I(l \in Z_i \wedge l \notin Y_i) + I(l \notin Z_i \wedge l \in Y_i)], \tag{6}$$

onde L é o conjunto de todas as etiquetas, é também uma métrica apropriada para avaliar problemas de classificação multi-etiqueta. Na próxima secção, os resultados de todas as métricas excepto a função de custo de Hamming serão apresentados na forma de percentagens.

Para verificar se as diferenças entre os resultados de duas abordagens são estatisticamente significativas, escolhemos aleatoriamente uma das execuções de cada uma das abordagens e aplicámos um teste binomial sobre a sua taxa de acerto, no caso das experiências sobre o Nível 1, e sobre a sua taxa de correspondência exacta, no caso das experiências sobre os Níveis 2 e 3. Ao longo da discussão apresentada na próxima secção, consideramos um nível de confiança de 95%, isto é, consideramos que existe uma diferença estatisticamente significativa entre duas abordagens se o valor-p do teste binomial for inferior a 0,05.

4 Resultados

Uma vez que cada nível da anotação hierárquica de actos de diálogo do corpus DIHANA tem características diferentes e coloca problemas de diferentes tipos, começamos por apresentar os resultados alcançados em cada um dos níveis de forma independente. Para além disso, como queremos avaliar a importância da informação de contexto dos níveis superiores, começamos no nível superior e descemos na hierarquia. Por fim, apresentamos os resultados obtidos na combinação hierárquica dos diferentes níveis.

4.1 Nível 1

Os resultados obtidos ao usar as duas abordagens de representação de segmento para prever as etiquetas de Nível 1 são mostrados na Tabela 4. Podemos ver que a abordagem baseada em CNNs tem melhor desempenho que a baseada em RNNs ($p \approx 0{,}04$). No entanto, ambas levam a uma taxa de acerto superior a 90% e a diferença entre elas é de apenas 0,5 pontos percentuais, o que sugere que a informação sobre intenção que são capazes de capturar é semelhante. No entanto, enquanto o treino da rede da abordagem baseada em CNNs demora em média 0,61 segundos por época e necessita de cerca de 27 épocas para convergir, o treino da rede da abordagem baseada em RNNs demora muito mais tempo, com uma média 17,63 segundos por época e 46 épocas para convergir.

Adicionalmente, tal como esperado, usar CNNs com janelas de contexto mais largas leva a melhores resultados ($p \approx 0{,}03$), o que confirma que as etiquetas genéricas do Nível 1 estão mais relacionadas com a estrutura do diálogo do que

Abordagem	**Acc** m	s
Recorrente (RNN)	91,20	0,06
Convolucional (CNN) w = [1,3]	91,46	0,12
Convolucional (CNN) w = [3,5]	91,70	0,13

Tabela 4: Taxa de acerto nas etiquetas de Nível 1 usando as duas abordagens de representação de segmento.

com palavras específicas. Ainda assim, uma vez que usamos três CNNs paralelas e existe sobreposição entre os dois conjuntos de janelas usados nas nossas experiências, a diferença em termos de taxa de acerto entre usar as janelas mais estreitas usadas por Liu et al. (2017) e as mais largas usadas por Kim (2014) é de apenas 0,24 pontos percentuais.

Relativamente à informação de contexto fornecida pelos segmentos anteriores, os resultados na Tabela 5 mostram que o primeiro segmento anterior é o mais importante, levando a uma melhoria da taxa de acerto na ordem dos 4,45 pontos percentuais ($p \approx 6{,}7e^{-167}$). Uma melhoria adicional de 1,77 pontos percentuais é alcançada fornecendo informação de dois segmentos adicionais ($p \approx 4{,}6e^{-58}$). Este padrão era esperado, uma vez que já tinha sido observado tanto no nosso estudo (Ribeiro et al., 2015) como no de Liu et al. (2017) no corpus SwDA, que também está anotado com etiquetas de actos de diálogo genéricas e independentes do domínio e da tarefa.

	Acc	
n	m	s
0	91,70	0,13
1	96,15	0,08
2	97,47	0,06
3	97,92	0,04

Tabela 5: Taxa de acerto nas etiquetas de Nível 1 usando informação de contexto de n segmentos anteriores.

Quando é utilizada informação de contexto de três segmentos anteriores, o classificador só falha a previsão de dois por cento dos segmentos. Este resultado tem em conta todos os segmentos do corpus DIHANA. No entanto, os segmentos do sistema são estruturados à priori e, portanto, são mais fáceis de prever do que os segmentos do utilizador. De facto, se considerarmos um cenário em que um sistema de diálogo tenta prever actos de diálogo, ele está ciente dos seus próprios actos e tem apenas de prever os dos seus interlocutores. Neste sentido, na Tabela 6 mostramos os resultados obtidos ao considerar os segmentos do utlizador e do sistema independentemente. Como esperado, a taxa de acerto média nos segmentos do sistema é de 99,91%. Nos segmentos do utilizador esse valor diminui para 95,17%, o que ainda assim revela um elevado desempenho.

Orador	**Acc** m	s
Utilizador	95.17	0.12
Sistema	99.91	0.00

Tabela 6: Taxa de acerto nas etiquetas de Nível 1 em segmentos do utilizador e do sistema.

Olhando para cada etiqueta individualmente, a mais difícil de identificar é a *Indefinida*, com uma sensibilidade de cerca de 57%. Isto era esperado, uma vez que essa etiqueta cobre todos os casos que não podem ser etiquetados com nenhuma das outras etiquetas, incluindo problemas no diálogo. Para todas as etiquetas restantes, o valor da sensibilidade é acima de 95%, sendo o mais baixo o da etiqueta *Resposta*, que é também aquela com valor mais baixo em termos de precisão (96%). Em ambos os casos, a confusão é tipicamente com a etiqueta *Pergunta*, o que faz sentido, uma vez que perguntas e respostas podem ter as mesmas palavras e diferir apenas em termos da sua ordem. De facto, se considerarmos questões em forma declarativa, pode não haver qualquer tipo de diferença.

Considerando estudos anteriores sobre reconhecimento de actos de diálogo no corpus DIHANA, só Tamarit & Martínez-Hinarejos (2008) é que avaliaram o desempenho no Nível 1 individualmente, atingindo uma taxa de acerto de 60,70%. No entanto, o estudo focava-se no uso de informação prosódica e, como tal, não é justo comparar os resultados com os nossos, pois a nossa abordagem tira partido das transcrições.

4.2 Level 2

Tal como referido na Secção 3.1, algumas etiquetas de Nível 1 só podem ser emparelhadas com a etiqueta *Nula* nos restantes níveis. Logo, os segmentos anotados com uma dessas etiquetas no Nível 1, ficam com as etiquetas dos restantes níveis definidas automaticamente, independentemente do seu conteúdo. Por isso, esses segmentos não são considerados nas nossas experiências sobre os Níveis 2 e 3.

De forma semelhante ao que observámos para o Nível 1, na Tabela 7 podemos ver que usar a abordagem de representação de segmento baseada em CNNs leva a melhores resultados do que

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
Abordagem	m	s	m	s	m	s	m	s	m	s	m	s
RNN	69,65	0,50	70,42	0,48	71,10	0,46	70,51	0,47	70,68	0,47	0,0381	0,0004
CNN w = [1,3]	70,71	0,33	71,58	0,33	72,30	0,33	71,74	0,34	71,87	0,33	0,0381	0,0002
CNN w = [3,5]	70,24	0,27	71,17	0,26	71,93	0,28	71,33	0,26	71,48	0,26	0,0383	0,0000

Tabela 7: Resultados obtidos no Nível 2 usando as duas abordagens de representação de segmento.

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
n	m	s	m	s	m	s	m	s	m	s	m	s
0	70,71	0,33	71,58	0,33	72,30	0,33	71,74	0,34	71,87	0,33	0,0381	0,0002
1	91,07	0,14	91,52	0,13	91,84	0,13	91,67	0,13	91,68	0,13	0,0121	0,0002
2	92,52	0,09	92,99	0,08	93,30	0,08	93,12	0,09	93,14	0,08	0,0101	0,0001
3	92,97	0,12	93,45	0,11	93,75	0,09	93,61	0,11	93,60	0,10	0,0094	0,0001

Tabela 8: Resultados obtidos no Nível 2 usando informação de contexto de n segmentos anteriores.

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
n	m	s	m	s	m	s	m	s	m	s	m	s
0	93,18	0,18	93,68	0,16	93,99	0,15	93,87	0,15	93,63	0,15	0,0092	0,0002
1	94,28	0,15	94,75	0,14	95,06	0,13	94,91	0,13	94,91	0,13	0,0077	0,0002
2	94,29	0,05	94,76	0,05	95,06	0,05	94,91	0,06	94,91	0,06	0,0077	0,0001
3	94,38	0,11	94,84	0,11	95,15	0,12	94,97	0,12	94,99	0,12	0,0075	0,0001

Tabela 9: Resultados obtidos no Nível 2 usando informação de Nível 1 de n segmentos anteriores.

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
Orador	m	s	m	s	m	s	m	s	m	s	m	s
Utilizador	91,28	0,24	92,08	0,21	92,62	0,19	92,32	0,22	92,34	0,21	0,0115	0,0003
Sistema	98,43	0,09	98,44	0,08	98,44	0,08	98,45	0,07	98,44	0,08	0,0024	0,0001

Tabela 10: Resultados obtidos no Nível 2 em segmentos do utilizador e do sistema.

a baseada em RNNs. A única excepção é o resultado ao nível da função de custo de Hamming que, em média, é igual para ambas as abordagens. Em todas as outras métricas, a abordagem baseada em CNNs supera a baseada em RNNs por mais de um ponto percentual ($p \approx 1{,}1e^{-14}$). No entanto, neste caso, a discrepância no número de épocas de treino necessárias para haver convergência é menor, com uma média de 46 para a abordagem baseada em CNNs e 56 para a baseada em RNNs. Para além disso, uma vez que são considerados menos segmentos, os tempos de treino por época são reduzidos para 0,40 e 11,67 segundos, respectivamente.

Por outro lado, contrariamente ao observado para o Nível 1, usar janelas de contexto mais estreitas parece levar a melhores resultados. No entanto, a diferença não é estatisticamente significativa ($p \approx 0{,}12$). Ainda assim, isto mostra que as etiquetas dependentes do domínio estão mais relacionadas com palavras específicas do que as etiquetas genéricas do Nível 1. Para além disso, uma vez que o número de etiquetas por segmento é tipicamente baixo, os classificadores tendem a evitar escolher etiquetas incorrectas, o que se reflecte numa precisão mais alta do que a sensibilidade em todas as abordagens.

Os resultados mostrados na Tabela 8 mostram que, de forma semelhante ao que acontece no Nível 1, os segmentos anteriores fornecem informação de contexto relevante para a tarefa. No entanto, neste caso, a importância do primeiro segmento anterior é mais pronunciada, levando a uma redução do custo para menos de um terço e melhorando as restantes métricas em cerca de 20 pontos percentuais ($p \approx 5{,}0e^{-324}$). Isto faz sentido, considerando que os diálogos incluem uma grande quantidade de pares pergunta-resposta focados no mesmo tipo de informação, que é o foco das etiquetas de Nível 2. Logo, nesses casos, as etiquetas de Nível 2 dos dois segmentos são as mesmas e, por isso, as etiquetas do

primeiro segmento anterior fornecem uma pista importante para a identificação das etiquetas do segmento actual.

Na Tabela 9, podemos ver que informação extraída do Nível 1 também é importante. Usar informação do segmento actual leva a uma melhoria que, embora significativa ($p \approx 0{,}01$), é apenas na ordem dos 0,2 pontos percentuais. No entanto, considerar também a etiqueta de Nível 1 do segmento anterior leva a uma melhoria na ordem dos 1,5 pontos percentuais ($p \approx 8{,}7e^{-6}$). Isto continua a ser explicado pela presença de um grande número de pares pergunta-resposta nos diálogos, uma vez que se o segmento anterior estiver etiquetado como *Pergunta* no Nível 1, então é provável que o segmento actual tenha as mesmas etiquetas de Nível 2 que esse segmento. Utilizar informação extraída de segmentos adicionais não leva a melhorias significativas ($p \approx 0{,}74$).

De forma semelhante ao que observámos para o Nível 1, o desempenho nos segmentos do sistema é diferente do nos segmentos do utilizador. Na Tabela 10 podemos ver que nos segmentos do sistema, os resultados ao nível de todas as métricas percentuais rondam os 98,4%, enquanto nos segmentos do utilizador a taxa de correspondência exacta é de 91,28% e as restantes métricas percentuais rondam os 92%.

Considerando as etiquetas individualmente, a melhor abordagem não é capaz de identificar nenhuma das três etiquetas menos predominantes no corpus. No entanto, isto era esperado, uma vez que nenhuma delas ocorre em mais de 29 segmentos. Como tal, elas são irrelevantes do ponto de vista de uma abordagem focada em reduzir o erro no corpus como um todo e necessitam de abordagens especializadas ou de mais dados para serem identificadas. O valor da medida-F para a etiqueta *Hora de Chegada* é de cerca de 75%, uma vez que esta é facilmente confundível com a etiqueta *Hora de Partida* e é a menos predominante das duas. Embora a precisão da etiqueta *Tipo de Comboio* seja acima de 95%, o valor da sensibilidade para a mesma etiqueta é de apenas 87%. Isto acontece porque a etiqueta aparece em apenas 2% dos segmentos. Como tal, em segmentos que se focam em múltiplos aspectos, a informação das palavras relacionadas com o tipo de comboio é descartada em favor de informação que permita identificar as etiquetas mais predominantes. Todas as etiquetas restantes têm um valor de medida-F acima dos 95% com balanço entre a precisão e a sensibilidade.

Os estudos anteriores sobre o reconhecimento de actos de diálogo no corpus DIHANA não exploraram o Nível 2 individualmente, mas sim em combinação com o Nível 1, usando a combinação das etiquetas dos dois níveis como conjunto de etiquetas e abordando o problema como um problema de classificação de etiqueta única semelhante ao colocado pelo Nível 1. Logo, os nossos resultados no Nível 2 não podem ser comparados directamente com os desses estudos. Os resultados obtidos na combinação dos dois níveis são discutidos na Secção 4.4.

4.3 Nível 3

A Tabela 11 mostra que, de forma semelhante ao que observámos nos restantes níveis, usar a abordagem de representação de segmento baseada em CNNs leva a melhores resultados do que a baseada em RNNs ($p \approx 9{,}6e^{-5}$). No entanto, neste caso a diferença é menos pronunciada. De facto, ao usar o conjunto de janelas de contexto mais largas, a abordagem baseada em CNNs tem pior desempenho que a baseada em RNNs ($p \approx 1{,}2e^{-4}$). Isto deve-se ao facto de o Nível 3 se focar na informação que é referida explicitamente nos segmentos e, como tal, ser ainda mais orientado a palavras específicas do que o Nível 2. Este facto também explica que, em média, os resultados de todas as métricas percentuais sejam superiores a 96%. Os tempos médios por época são iguais aos registados para o Nível 2. No entanto, são necessárias mais épocas para atingir convergência — 86 para a abordagem baseada em RNNs e 80 para a baseada em CNNs.

Os resultados da Tabela 12 mostram que, neste caso, a melhoria obtida ao usar informação dos segmentos anteriores ao mesmo nível é desprezável e não é estatisticamente significativa ($p \approx 0{,}48$). Uma vez mais, isto pode ser explicado pela natureza do Nível 3 e o seu foco no que é referido explicitamente no segmento actual. Logo, informação relativa ao que é referido explicitamente nos segmentos anteriores não é relevante.

Na Tabela 13 podemos ver que a informação fornecida pelo Nível 2 é ligeiramente superior à fornecida pelas etiquetas de Nível 3 dos segmentos anteriores. Neste caso, considerar a etiqueta de Nível 2 do mesmo segmento leva uma melhoria estatisticamente significativa ($p \approx 0{,}03$). Esta melhoria pode ser explicada pelo facto de que quando um tipo de informação é referido explicitamente num segmento, tipicamente este também é focado pelo mesmo segmento e, por isso, é comum haver sobreposição das etiquetas dos Níveis 2 e 3. Considerar informação de Nível 2 de segmentos adicionais não leva a melhorias estatisticamente significativas ($p \approx 0{,}13$).

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
Abordagem	m	s	m	s	m	s	m	s	m	s	m	s
RNN	95,79	0,24	96,61	0,29	96,84	0,29	96,81	0,30	96,78	0,30	0,0043	0,0004
CNN w = [1,3]	96,01	0,08	96,88	0,10	97,11	0,10	97,08	0,12	97,05	0,11	0,0040	0,0000
CNN w = [3,5]	95,35	0,23	96,26	0,18	96,51	0,17	96,45	0,15	96,44	0,16	0,0046	0,0002

Tabela 11: Resultados obtidos no Nível 3 usando as duas abordagens de representação de segmento.

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
n	m	s	m	s	m	s	m	s	m	s	m	s
0	96,01	0,08	96,88	0,10	97,11	0,10	97,08	0,12	97,05	0,11	0,0040	0,0000
1	96,05	0,13	96,91	0,10	97,14	0,09	97,10	0,09	97,08	0,10	0,0039	0,0001
2	96,10	0,16	96,95	0,11	97,17	0,11	97,14	0,10	97,12	0,10	0,0039	0,0002
3	96,10	0,16	96,96	0,13	97,19	0,13	97,14	0,11	97,13	0,12	0,0039	0,0001

Tabela 12: Resultados obtidos no Nível 3 usando informação de contexto de n segmentos anteriores.

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
n	m	s	m	s	m	s	m	s	m	s	m	s
0	96,20	0,15	97,03	0,11	97,25	0,11	97,20	0,10	97,19	0,11	0,0037	0,0002
1	96,24	0,09	97,05	0,08	97,28	0,08	97,22	0,07	97,21	0,08	0,0037	0,0000
2	96,29	0,08	97,11	0,08	97,34	0,09	97,27	0,09	97,26	0,09	0,0036	0,0000
3	96,17	0,06	97,00	0,06	97,23	0,06	97,18	0,06	97,17	0,06	0,0038	0,0000

Tabela 13: Resultados obtidos no Nível 3 usando informação de Nível 2 de n segmentos anteriores.

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
n	m	s	m	s	m	s	m	s	m	s	m	s
0	96,32	0,12	97,13	0,10	97,36	0,09	97,29	0,10	97,28	0,09	0,0036	0,0001
1	96,29	0,14	97,10	0,12	97,33	0,12	97,26	0,11	97,26	0,12	0,0037	0,0001
2	96,34	0,12	97,14	0,11	97,36	0,10	97,30	0,11	97,30	0,11	0,0036	0,0001
3	96,30	0,13	97,13	0,11	97,35	0,10	97,31	0,10	97,29	0,10	0,0037	0,0001

Tabela 14: Resultados obtidos no Nível 3 usando informação de Nível 1 de n segmentos anteriores.

	MR		**Acc**		**P**		**R**		**F_1**		**HL**	
Orador	m	s	m	s	m	s	m	s	m	s	m	s
Utilizador	95,58	0,16	95,62	0,17	95,62	0,17	95,65	0,18	95,63	0,17	0,0044	0,0002
Sistema	97,55	0,12	99,06	0,06	99,52	0,06	99,25	0,03	99,33	0,04	0,0024	0,0001

Tabela 15: Resultados obtidos no Nível 3 em segmentos do utilizador e do sistema.

Uma vez que as etiquetas de Nível 1 estão relacionadas com a intenção genérica por trás do segmento, elas não têm relação directa com que é explicitamente referido no segmento e, portanto, com as etiquetas de Nível 3. Isto é confirmado pelos resultados da Tabela 14, que mostram que a melhoria obtida ao utilizar informação do Nível 1 é desprezável e não é estatisticamente significativa ($p \approx 0{,}13$).

Na Tabela 15, podemos ver que, neste caso, a diferença de desempenho nos segmentos do utilizador e do sistema não é tão pronunciada. Mais uma vez, isso é explicado pelo facto de que o Nível 3 é altamente focado em palavras específicas e, portanto, o facto de os segmentos do sistema serem estruturados à priori não tem a mesma influência na classificação.

Considerando as etiquetas individualmente, à semelhança do que acontece no Nível 2, a melhor abordagem é incapaz de identificar as etiquetas menos predominantes, *Duração* e *Serviço*, uma vez que nenhuma delas aparece em mais de 19 segmentos. Das restantes, a etiqueta *Hora de Chegada* é aquela com menor valor de

sensibilidade, 88%, uma vez que é facilmente confundível com a mais predominante *Hora de Partida*. Todas as etiquetas restantes têm um valor de medida-F acima de 97% com balanço entre precisão e sensibilidade.

Tal como o Nível 2, os estudos anteriores sobre o reconhecimento de actos de diálogo no corpus DIHANA não exploraram o Nível 3 individualmente, mas sim em combinação com os restantes níveis. Consequentemente, também não é possível comparar directamente os nossos resultados no Nível 3 com os desses estudos. A combinação hierárquica dos vários níveis é explorada na próxima secção.

4.4 Classificação Hierárquica

Tal como referido anteriormente, os estudos anteriores sobre reconhecimento de actos de diálogo no corpus DIHANA não exploraram os níveis específicos da tarefa independentemente, mas sim em combinação com os níveis acima. Isto faz sentido dum ponto de vista hierárquico, uma vez que, supostamente, cada nível é dependente dos que estão acima dele. No entanto, tal como discutido na Secção 3.1, uma vez que cada nível se foca num aspecto diferente relativo à intenção do orador, a única restrição imposta pelo esquema de anotação é que segmentos anotados com uma etiqueta de Nivel 1 relacionada com a estrutura do diálogo ou problemas de comunicação não podem ter etiquetas nos restantes níveis. Ainda assim, os resultados reportados nas secções anteriores mostram que a capacidade de prever as etiquetas de um determinado nível aumenta quando é usada informação de contexto extraída do nível directamente acima. Para além disso, para identificar correctamente a intenção do seu interlocutor, um sistema de diálogo tem de ser capaz de prever correctamente as etiquetas dos três níveis em conjunto. Por isso, também avaliamos o desempenho das nossas abordagens na combinação hierárquica dos vários níveis.

Os estudos anteriores sobre a tarefa abordaram o problema da classificação combinada dos diferentes níveis como um problema de classificação de etiqueta única, em que cada combinação de etiquetas presentes no corpus é considerada uma única etiqueta independente. No entanto, esta abordagem apresenta duas falhas. Por um lado, trata-se de uma simplificação do problema, na medida em que limita as etiquetas possíveis às combinações existentes no corpus. Por outro lado, não tem em conta a natureza multi-etiqueta dos níveis específicos da tarefa.

Contrariamente a esses estudos, nós abordamos o problema hierarquicamente, combinando os melhores classificadores para cada nível. Ou seja, para cada segmento, começamos por prever a sua etiqueta de Nível 1 usando o classificador baseado em CNNs com janelas de contexto mais largas e informação de contexto extraída de três segmentos anteriores. Em seguida, prevemos as suas etiquetas de Nível 2 usando o classificador baseado em CNNs com janelas de contexto mais estreitas, informação de contexto de Nível 2 de três segmentos anteriores e informação de contexto de Nível 1 do segmento actual e do anterior. Por fim, prevemos as suas etiquetas de Nível 3 usando o classificador baseado em CNNs com janelas de contexto mais estreitas e informação de contexto de Nível 2 extraída do segmento actual. De forma a ter em conta o facto de os classificadores dos Níveis 2 e 3 não terem sido treinados nos segmentos com etiquetas de Nível 1 que não permitem etiquetas nos restantes níveis, se o classificador de Nível 1 prevê uma dessas etiquetas para o segmento, os restantes níveis são automaticamente classificados como não tendo etiquetas.

Usando esta abordagem hierárquica, os níveis inferiores ainda são considerados problemas de classificação multi-etiqueta. Logo, todas as combinações de etiquetas são possíveis e não apenas aquelas que aparecem no corpus. Ainda assim, para confirmar que o problema abordado pelos estudos anteriores é realmente mais simples, apresentamos também os resultados alcançados quando a tarefa é abordada como um problema de classificação de etiqueta única. Para obter esses resultados, usámos um classificador com a mesma arquitectura que o melhor classificador de Nível 1, ou seja, um classificador baseado em CNNs com janelas de contexto mais largas e informação de contexto extraída de três segmentos anteriores. No entanto, neste caso, o classificador foi treinado para prever a combinação de todas as etiquetas do segmento de uma só vez e cada uma dessas combinações é vista como uma etiqueta independente.

Para comparação com os resultados obtidos em estudos anteriores, utilizamos a taxa de correspondência exata para avaliar o desempenho quer da abordagem hierárquica, quer da de etiqueta única. Portanto, se a previsão da etiqueta de Nível 1 for incorrecta ou se houver alguma etiqueta de Nível 2 ou 3 em falta ou adicional, toda a previsão para o segmento é considerada errada.

A Tabela 16 mostra os resultados obtidos na combinação dos Níveis 1 e 2. Usando a abordagem hierárquica, obtivemos, em média, 94,28% de taxa de correspondência exacta, um resultado

já acima dos 93,40% reportados por Martínez-Hinarejos et al. (2008) ($p \approx 3{,}0e^{-8}$) e em linha com os 94,08% reportados por Gambäck et al. (2011) ($p \approx 0{,}20$). Ao abordar a tarefa como um problema de classificação de etiqueta única, obtivemos 96,24% de taxa de correspondência exacta, um resultado que é quase dois pontos percentuais acima do resultado obtido usando a abordagem hierárquica ($p \approx 7{,}0e^{-43}$). Isto confirma que a visão do problema como um problema de classificação de etiqueta única é realmente uma simplificação.

	MR	
Abordagem	m	s
Hierárquica	94,28	0,03
Etiqueta Única	96,24	0,06
Martínez-Hinarejos et al. (2008)	93,40	
Gambäck et al. (2011)	94,08	

Tabela 16: Resultos obtidos na combinação dos Níveis 1 e 2.

A Tabela 17 mostra os resultados obtidos na combinação dos três níveis. Podemos ver que a maioria das conclusões tiradas para a combinação dos Níveis 1 e 2 também pode ser tirada neste caso. Usando a abordagem hierárquica, obtivemos, em média, uma taxa de correspondência exacta de 92,34 %, que está acima dos 89,70% reportados por Martínez-Hinarejos et al. (2008) e dos 90,97% reportados por Gambäck et al. (2011). No entanto, enquanto na combinação dos dois níveis superiores o resultado da abordagem hierárquica não é estatisticamente diferente do reportado por Gambäck et al. (2011), neste caso há uma melhoria estatisticamente significativa de 1,37 pontos percentuais ($p \approx 6{,}6e^{-14}$). Ao abordar a tarefa como um problema de classificação de etiqueta única, a taxa de correspondência exacta é melhorada para 93,98% ($p \approx 1{,}5e^{-22}$), confirmando uma vez mais que o problema é mais simples.

	MR	
Abordagem	m	s
Hierárquica	92,34	0,04
Etiqueta Única	93,98	0,19
Martínez-Hinarejos et al. (2008)	89,70	
Gambäck et al. (2011)	90,97	

Tabela 17: Resultos obtidos na combinação de todos os níveis.

5 Conclusões

Neste artigo explorámos o reconhecimento automático de actos de diálogo no corpus DIHANA. Este corpus e o seu esquema de anotação em três níveis colocam problemas que não têm sido explorados desde que os estudos sobre o reconhecimento de actos de diálogo começaram a focar-se em dados em inglês e, especialmente, no corpus SwDA. O primeiro problema diz respeito à diferença de língua, uma vez que o espanhol tem características diferentes do inglês. Adicionalmente, ao contrário do problema de classificação plana e de etiqueta única colocado pelas anotações SWBD-DAMSL do corpus SwDA, as anotações de actos de diálogo do corpus DIHANA colocam um problema de classificação hierárquica. Para além disso, os dois níveis inferiores dessa hierarquia colocam problemas de classificação multi-etiqueta. Por isso, estudámos como as melhores abordagens para o reconhecimento de actos de diálogo em dados em inglês podem ser aplicadas a estes problemas e quais os aspectos dessas abordagens que são relevantes para a previsão das etiquetas de cada nível, de acordo com suas características.

Uma conclusão comum a todos os níveis é que a abordagem de representação do segmento baseada em CNNs leva a um melhor desempenho do que a baseada em RNNs. Esta abordagem, aplicada ao reconhecimento de actos de diálogo em inglês por Liu et al. (2017), apresenta três CNNs temporais paralelas com janelas de contexto de diferentes tamanhos. Assim, a abordagem de representação do segmento tem em conta conjuntos de palavras de diferentes tamanhos e, dependendo dos tamanhos das janelas, é capaz de capturar informação referente quer a palavras específicas, quer à estrutura do segmento. As etiquetas genéricas e independentes da tarefa do Nível 1 estão mais relacionadas com a estrutura do segmento pelo que os melhores resultados foram obtidos utilizando janelas de contexto mais largas. Por outro lado, as etiquetas específicas da tarefas dos Níveis 2 e 3 estão mais relacionadas com palavras específicas e, por isso, a utilização de um conjunto de janelas mais estreitas levou a um melhor desempenho. Seleccionar um conjunto de janelas adequado é especialmente relevante para prever as etiquetas de Nível 3, uma vez que, ao usar janelas mais largas, a abordagem com CNNs teve um desempenho pior do que a baseada em RNNs. Isso é explicável pela natureza desse nível, que se foca no tipo de informação que é explicitamente referido nos segmentos e, portanto, a classificação de um segmento é dada pela presença de palavras específicas.

A relação entre as etiquetas de Nível 3 e a presença de palavras específicas no segmento explica também o facto de a informação de contexto extraída dos segmentos anteriores não ser relevante para a previsão dessas etiquetas. Por outro lado, essa informação é relevante para prever as etiquetas dos restantes níveis. No Nível 1, as experiências revelaram um padrão semelhante ao observado tanto no nosso estudo anterior (Ribeiro et al., 2015), como no de Liu et al. (2017) no corpus SwDA, que também está anotado com etiquetas genéricas e independentes da tarefa. No entanto, a importância da informação de contexto dos segmentos anteriores foi especialmente pronunciada nas experiências sobre o Nível 2, reduzindo o valor da função de custo de Hamming para menos de um terço e melhorando as restantes métricas em mais de 20 pontos percentuais. O Nível 2 foca-se no tipo de informação implicitamente focada pelo segmento. Logo, uma vez que os diálogos no corpus DIHANA apresentam múltiplos pares de segmentos focados no mesmo tipo de informação, os segmentos anteriores, especialmente o primeiro, fornecem uma pista importante para a classificação do segmento actual.

Ainda considerando o Nível 2 e as características dos diálogos, a maioria dos pares de segmentos focados no mesmo tipo de informação são pares pergunta-resposta. *Pergunta* e *Resposta* são etiquetas de Nível 1. Por isso, a informação de contexto de Nível 1 extraída quer do segmento actual, quer dos anteriores, também fornece pistas para a previsão de etiquetas de Nível 2. Por outro lado, essa informação é irrelevante para prever etiquetas de Nível 3. No entanto, existe uma relação entre o tipo de informação que é implicitamente focada num segmento e aquela que é explicitamente mencionada nele. Logo, tipicamente, existe sobreposição entre os conjuntos de etiquetas de Nível 2 e 3 de um segmento. Consequentemente, usar informação de contexto extraída do Nível 2 leva a ligeiras melhorias no desempenho ao prever etiquetas de Nível 3.

No corpus DIHANA, os segmentos do sistema são estruturados à priori e, portanto, os seus actos de diálogo são mais fáceis de prever do que os dos segmentos do utilizador. Para além disso, um sistema de diálogo está ciente dos seus próprios actos de diálogo e tem apenas de prever os dos segmentos dos seus interlocutores. Logo, nesse cenário, apenas o desempenho nos segmentos do utilizador é relevante. Como esperado, o desempenho foi mais elevado nos segmentos do sistema em todos os níveis. No entanto, nos segmentos do utilizador, a taxa de acerto no Nível 1 e a taxa de correspondência exacta nos restantes níveis ainda ficaram acima de 90%. Para além disso, é importante referir que, como o Nível 3 é altamente relacionado com palavras específicas, a diferença de desempenho não é tão pronunciada nesse nível.

Por fim, ao combinar hierarquicamente os melhores classificadores para cada nível, obtivemos, em média, uma taxa de correspondência exacta de 94,28% na combinação dos Níveis 1 e 2 e 92,34% na combinação dos três níveis. Esses resultados são já em linha com ou superiores aos obtidos em estudos anteriores sobre o reconhecimento de actos de diálogo no corpus DIHANA. No entanto, esses estudos consideraram uma versão simplificada do problema, reduzindo-o a um problema de classificação de etiqueta única, em que a etiqueta de um segmento consiste na concatenação das etiquetas dos três níveis. Uma vez que esta abordagem considera apenas as combinações de etiquetas presentes no corpus, o número de etiquetas possíveis é reduzido em comparação com a nossa abordagem, que aborda a previsão de etiquetas dos Níveis 2 e 3 como problemas de classificação multi-etiqueta. Ao abordar o problema de forma comparável à desses estudos, os valores anteriores aumentam para 96,24% e 93,98%, respectivamente.

Como trabalho futuro, seria interessante avaliar se as conclusões tiradas deste estudo sobre dados em espanhóis e, anteriormente, sobre dados em inglês, se mantêm para dados em outras línguas com tipologia morfológica diferente. Em termos de reconhecimento de actos de diálogo multi-etiqueta, seria interessante explorar o uso de outras funções de custo ao treinar a rede, especialmente uma baseada na medida-F, que não é tão influenciada pelo número reduzido de classes positivas por segmento como a função de custo de Hamming. Para além disso, é importante avaliar se as abordagens de representação do segmento baseadas em *tokenização* ao nível do carácter são capazes de capturar informação adicional para prever as etiquetas específicas da tarefa. Também seria interessante explorar meios para realizar a classificação hierárquica dos múltiplos níveis usando uma única rede em vez de três classificadores independentes. Por fim, é importante avaliar a deterioração do desempenho num cenário real. Ou seja, um em que o sistema de diálogo não é simulado e, portanto, tem de lidar com problemas relacionados com o Reconhecimento Automático de Fala (ASR) e usar etiquetas previstas automaticamente como informação de contexto.

Agradecimentos

Este estudo foi financiado por fundos nacionais, através da Fundação para a Ciência e a Tecnologia (FCT), com a referência UID/CEC/50021/2019, e pela Universidade de Lisboa.

Referências

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu & Xiaoqiang Zheng. 2016. TensorFlow: large-scale machine learning on heterogeneous systems. `arXiv:1603.04467v2`.

Alcácer, N., J. M. Benedí, F. Blat, R. Granell, C. D. Martínez & F. Torres. 2005. Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. Em *10th International Conference on Speech and Computer (SPECOM)*, 583–586.

Alexandersson, Jan, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz & Melanie Siegel. 1998. Dialogue Acts in VERBMOBIL-2 Second Edition. Relatório técnico. DFKI.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller et al. 1991. The HCRC Map Task Corpus. *Language and Speech* 34(4). 351–366. 10.1177/002383099103400404.

Ang, Jeremy, Yang Liu & Elizabeth Shriberg. 2005. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. Em *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1061–1064. 10.1109/ICASSP.2005.1415300.

Benedí, José-Miguel, Eduardo Lleida, Amparo Varona, María-José Castro, Isabel Galiano, Raquel Justo, Iñigo López de Letona & Antonio Miguel. 2006. Design and Acquisition of a Telephone Spontaneous Speech Dialogue Corpus in Spanish: DIHANA. Em *Language, Resources and Evaluation Conference (LREC)*, 1636–1639.

Bunt, Harry, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis & David R. Traum. 2012. ISO 24617-2: A Semantically-Based Standard for Dialogue Annotation. Em *Language Resources and Evaluation Conference (LREC)*, 430–437.

Bunt, Harry, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven & Alex Fang. 2016. The DialogBank. Em *Language Resources and Evaluation Conference (LREC)*, 3151–3158.

Cardellino, Cristian. 2016. Spanish Billion Word Corpus and Embeddings. https://crscardellino.github.io/SBWCE/.

Carletta, Jean, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma & Pierre Wellner. 2005. The AMI Meeting Corpus: A Pre-Announcement. Em *Machine Learning for Multimodal Interaction (MLMI)*, 28–39. 10.1007/11677482_3.

Carroll, John M. & Michael K. Tanenhaus. 1978. Functional Clauses and Sentence Segmentation. *Journal of Speech, Language, and Hearing Research* 21(4). 793–808. 10.1044/jshr.2104.793.

Chollet, François et al. 2015. Keras: The Python Deep Learning Library. https://keras.io/.

Conneau, Alexis, Holger Schwenk, Loïc Barrault & Yann Lecun. 2017. Very Deep Convolutional Networks for Text Classification. Em *15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 1107–1116.

Coria, Sergio R. & Luis A. Pineda. 2005. Predicting Obligation Dialogue Acts from Prosodic and Speaker Information. *Research in Computing Science* 14. 137–148.

Coria, Sergio R. & Luis A. Pineda. 2006. Predicting Dialogue Acts from Prosodic Information. Em *Computational Linguistics and Intelligent Text Processing*, 355–365. 10.1007/11671299_37.

Coria, Sergio R. & Luis A. Pineda. 2009. An Analysis of Prosodic Information for the Recognition of Dialogue Acts in a Multimodal Corpus in Mexican Spanish. *Computer Speech & Language* 23(3). 277–310. 10.1016/j.csl.2008.06.003.

Costantini, Erica, Susanne Burger & Fabio Pianesi. 2002. NESPOLE!'s Multilingual and Multimodal Corpus. Em *Language Resources and Evaluation Conference (LREC)*, 165–170.

Di Eugenio, Barbara, Zhuli Xie & Riccardo Serafin. 2010. Dialogue Act Classification, Higher Order Dialogue Structure, and Instance-Based Learning. *Dialogue and Discourse* 1(2). 81–104. 10.5087/dad.2010.002.

Díez, Jorge, Oscar Luaces, Juan José del Coz & Antonio Bahamonde. 2015. Optimizing Different Loss Functions in Multilabel Classifications. *Progress in Artificial Intelligence* 3(2). 107–118. 10.1007/s13748-014-0060-7.

Gambäck, Björn, Fredrik Olsson & Oscar Täckström. 2011. Active Learning for Dialogue Act Classification. Em *International Speech Communication Association (INTERSPEECH)*, 1329–1332.

Jekat, Susanne, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast & J. Joachim Quantz. 1995. Dialogue Acts in VERBMOBIL. Relatório técnico. DFKI.

Ji, Yangfeng, Gholamreza Haffari & Jacob Eisenstein. 2016. A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 332–342. 10.18653/v1/N16-1037.

Jurafsky, Dan, Elizabeth Shriberg & Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. Relatório Técnico. Draft 13 University of Colorado, Institute of Cognitive Science.

Kalchbrenner, Nal & Phil Blunsom. 2013. Recurrent Convolutional Neural Networks for Discourse Compositionality. Em *Workshop on Continuous Vector Space Models and their Compositionality*, 119–126.

Kay, Martin, Peter Norvig & Mark Gawron. 1992. *VERBMOBIL: A Translation System for Face-to-Face Dialog*. University of Chicago Press.

Khanpour, Hamed, Nishitha Guntakandla & Rodney Nielsen. 2016. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. Em *26th International Conference on Computational Linguistics (COLING)*, 2012–2021.

Kim, Seokhwan, Luis Fernando D'Haro, Rafael E. Banchs, Jason Williams & Matthew Henderson. 2017. The Fourth Dialog State Tracking Challenge. Em *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, 435–449. 10.1007/978-981-10-2585-3_36.

Kim, Yoon. 2014. Convolutional Neural Networks for Sentence Classification. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. 10.3115/v1/D14-1181.

Kingma, Diederik P. & Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. Em *ICLR*, http://arxiv.org/abs/1412.6980.

Král, Pavel & Christophe Cerisara. 2010. Dialogue Act Recognition Approaches. *Computing and Informatics* 29(2). 227–250.

Lee, Ji Young & Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 515–520. 10.18653/v1/N16-1062.

Levin, Lori, Ann Thymé-Gobbel, Alon Lavie, Klaus Ries & Klaus Zechner. 1998. A Discourse Coding Scheme for Conversational Spanish. Em *5th International Conference on Spoken Language Processing (ICSLP)*, paper 1000.

Levin, Lori S., Klaus Ries, Ann Thymé-Gobbel & Alon Lavie. 1999. Tagging Of Speech Acts And Dialogue Games In Spanish Call Home. Em *Workshop On Towards Standards And Tools For Discourse Tagging*, 42–47.

Liu, Yang, Kun Han, Zhao Tan & Yun Lei. 2017. Using Context Information for Dialog Act Classification in DNN Framework. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2160–2168. 10.18653/v1/D17-1231.

Margolis, Anna, Karen Livescu & Mari Ostendorf. 2010. Domain Adaptation with Unlabeled Data for Dialog Act Tagging. Em *Workshop on Domain Adaptation for Natural Language Processing (DANLP)*, 45–52.

Martínez-Hinarejos, Carlos D., José-Miguel Benedí & Ramón Granell. 2008. Statistical Framework for a Spanish Spoken Dialogue Corpus. *Speech Communication* 50(11–12). 992–1008. 10.1016/j.specom.2008.05.011.

Martínez-Hinarejos, Carlos D., José-Miguel Benedí & Vicent Tamarit. 2015. Unsegmented Dialogue Act Annotation and Decoding with N-Gram Transducers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(1). 198–211. 10.1109/TASLP.2014.2377595.

Martínez-Hinarejos, Carlos D., Emilio Sanchis, Fernando García-Granada & Pablo Aibar. 2002. A Labelling Proposal to Annotate Dialogues. Em *Language Resources and Evaluation Conference (LREC)*, vol. 5, 1566–1582.

Mezza, Stefano, Alessandra Cervone, Evgeny A. Stepanov, Giuliano Tortoreto & Giuseppe Riccardi. 2018. ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents. Em *International Conference on Computational Linguistics (COLING)*, 3539–3551.

Mikolov, Tomas, Martin Karafiát, Lukás Burget, Jan Cernocký & Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. Em *International Speech Communication Association (INTERSPEECH)*, 1045–1048.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. Em *Conference on Neural Information Processing Systems (NIPS)*, 3111–3119.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Petukhova, Volha, Martin Gropp, Dietrich Klakow, Anna Schmidt, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, Olivier Deroo, Ronny Egeler, Uwe Meinz & Steffen Liersch. 2014. The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. Em *Language Resources and Evaluation Conference (LREC)*, 252–258.

Ribeiro, Eugénio, Ricardo Ribeiro & David Martins de Matos. 2015. The Influence of Context on Dialogue Act Recognition. *CoRR* abs/1506.00839. http://arxiv.org/abs/1506.00839.

Ribeiro, Eugénio, Ricardo Ribeiro & David Martins de Matos. 2016. Mapping the Dialog Act Annotations of the LEGO Corpus into the Communicative Functions of ISO 24617-2. *CoRR* abs/1612.01404. http://arxiv.org/abs/1612.01404.

Ribeiro, Eugénio, Ricardo Ribeiro & David Martins de Matos. 2018. A Study on Dialog Act Recognition using Character-Level Tokenization. Em *Artificial Intelligence: Methodology, Systems, and Applications (AIMSA)*, 93–103. 10.1007/978-3-319-99344-7_9.

Ries, Klaus. 1999. HMM and Neural Network Based Speech Act Detection. Em *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 497–500. 10.1109/ICASSP.1999.758171.

Schmitt, Alexander, Stefan Ultes & Wolfgang Minker. 2012. A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let's Go Bus Information System. Em *Language Resources and Evaluation Conference (LREC)*, 3369–3373.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Serafin, Riccardo & Barbara Di Eugenio. 2004. FLSA: Extending Latent Semantic Analysis with Features for Dialogue Act Classification. Em *42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 692–699. 10.3115/1218955.1219043.

Serafin, Riccardo, Barbara Di Eugenio & Michael Glass. 2003. Latent Semantic Analysis for Dialogue Act Classification. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, 94–96. 10.3115/1073483.1073515.

Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang & Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. Em *5th SIGdial Workshop on Discourse and Dialogue (SIGDIAL)*, 97–100.

Sorower, Mohammad S. 2010. A Literature Survey on Algorithms for Multi-Label Learning. Relatório técnico. Oregon State University.

Stolcke, Andreas, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin & Marie Meteer. 2000.

Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26(3). 339–373. 10.1162/089120100561737.

Tamarit, Vicent & Carlos D. Martínez-Hinarejos. 2008. Dialog Act Labeling in the DIHANA Corpus using Prosody Information. Em *V Jornadas en Tecnología del Habla*, 183–186.

Tran, Quan Hung, Ingrid Zukerman & Gholamreza Haffari. 2017a. A Generative Attentional Neural Network Model for Dialogue Act Classification. Em *55th Annual Meeting of the Association for Computational Linguistics*, vol. 2, 524–529. 10.18653/v1/P17-2083.

Tran, Quan Hung, Ingrid Zukerman & Gholamreza Haffari. 2017b. A Hierarchical Neural Model for Learning Sequences of Dialogue Acts. Em *15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 428–437.

Tran, Quan Hung, Ingrid Zukerman & Gholamreza Haffari. 2017c. Preserving Distributional Information in Dialogue Act Classification. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2151–2156. 10.18653/v1/D17-1229.

Villaseñor, Luis, Antonio Massé & Luis A. Pineda. 2001. The DIME Corpus. Em *Encuentro Internacional de Ciencias de la Computación*, vol. 2, 1–10.

Proposta recebida em Outubro 2018 e aceite para publicação em Maio 2019.

Avaliando Atributos para a Classificação de Estrutura Retórica em Resumos Científicos

Evaluating features for rhetorical structure classification in scientific abstracts

Alessandra Harumi Iriguti
Universidade Estadual de Maringá
alehairi@gmail.com

Valéria Delisandra Feltrim
Universidade Estadual de Maringá
vdfeltrim@uem.br

Resumo

A classificação de estrutura retórica é uma tarefa de PLN na qual se busca identificar os componentes retóricos de um discurso e seus relacionamentos. No caso deste trabalho, buscou-se identificar automaticamente categorias em nível de sentenças que compõem a estrutura retórica de resumos científicos. Especificamente, o objetivo foi avaliar o impacto de diferentes conjuntos de atributos na implementação de classificadores retóricos para resumos científicos escritos em português. Para isso, foram utilizados atributos superficiais (extraídos como valores TF-IDF e selecionados com o teste χ^2), atributos morfossintáticos (implementados pelo classificador AZPort) e atributos extraídos a partir de modelos de *word embeddings* (Word2Vec, Wang2Vec e GloVe, todos previamente treinados). Tais conjuntos de atributos, bem como as suas combinações, foram usados para o treinamento de classificadores usando os seguintes algoritmos de aprendizado supervisionado: *Support Vector Machines*, *Naive Bayes*, *K-Nearest Neighbors*, *Decision Trees* e *Conditional Random Fields* (CRF). Os classificadores foram avaliados por meio de validação cruzada sobre três *corpora* compostos por resumos de teses e dissertações. O melhor resultado, 94% de F1, foi obtido pelo classificador CRF com as seguintes combinações de atributos: (i) Wang2Vec–*Skip-gram* de dimensões 100 com os atributos provenientes do AZPort; (ii) Wang2Vec–Skip-gram e GloVe de dimensão 300 com os atributos do AZPort; (iii) TF-IDF, AZPort e *embeddings* extraídos com os modelos Wang2Vec–*Skip-gram* de dimensões 100 e 300 e GloVe de dimensão 300. A partir dos resultados obtidos, conclui-se que os atributos provenientes do classificador AZPort foram fundamentais para o bom desempenho do classificador CRF, enquanto que a combinação com *word embeddings* se mostrou válida para a melhoria dos resultados.

Palavras chave

processamento de língua natural, classificação de estrutura retórica, resumos científicos em português

Abstract

Rhetorical structure classification is a NLP task in which we want to identify the rhetorical components of a discourse and its relationships. In this work, we aimed at automatically identifying categories at the sentential level that make up the rhetorical structure of scientific abstracts. Specifically, the purpose was to evaluate the impact of different sets of attributes on the implementation of rhetorical classifiers for scientific abstracts written in Portuguese. For this, we used superficial features (extracted as TF-IDF values and selected with the χ^2 test), morphosyntactic features (implemented by the AZPort classifier) and features extracted from word embeddings models (Word2Vec, Wang2Vec and GloVe, all of them previously trained). These sets of features, as well as their combinations, were used to train the following supervised learning classifiers: Support Vector Machines, Naive Bayes, K-Nearest Neighbors, Decision Trees and Conditional Random Fields (CRF). We evaluated the classifiers through cross-validation on three *corpora* composed by abstracts of theses and dissertations. The best result, 94% of F1, was obtained by the CRF classifier with the following combinations of features: (i) Wang2Vec–Skip-gram of 100 dimension with features from AZPort; (ii) Wang2Vec–Skip-gram and GloVe of 300 dimension with AZPort features; (iii) TF-IDF, AZPort features and embeddings extracted by Wang2Vec–Skip-gram model with dimensions of 100 and 300, and by GloVe model of dimension 300. From the results, we concluded that the AZPort features were fundamental for the performance of the CRF classifier, while the combination with word embeddings proved valid for improving the results.

Keywords

natural language processing, rhetorical structure classification, scientific abstracts in Portuguese

DOI: 10.21814/lm.11.1.273
This work is Licensed under a Creative Commons Attribution 4.0 License

1 Introdução

A classificação de estrutura retórica é uma tarefa de Processamento de Língua Natural (PLN) em que se busca identificar os componentes retóricos de um discurso e seus relacionamentos. São essas relações que definem como o conteúdo apresentado está relacionado entre si e como cada parte do texto contribui para satisfazer os objetivos e intenções do autor (Romeiro, 2016). Os componentes retóricos podem ser analisados com uma granularidade mais fina, como no caso da Teoria de Estrutura Retórica (Rhetorical Structure Theory - RST) (Mann & Thompson, 1987, 1988), que identifica relações entre segmentos que podem, por exemplo, compor uma mesma sentença, ou com uma granularidade mais grossa, em que os componentes retóricos são blocos de uma ou mais sentenças que juntos revelam a macro-estrutura de um texto.

A organização retórica em nível de macro-estrutura tem sido especialmente investigada no contexto dos textos científicos. Do ponto de vista linguístico, esses estudos buscam identificar modelos de estrutura retórica que caracterizem os movimentos retóricos observados nas diferentes seções desses textos. Weissberg & Buker (1990), Booth et al. (2005) e Swales & Feak (1994) são exemplos de autores que investigaram estruturas retóricas específicas do gênero científico. Booth et al. (2005) e Weissberg & Buker (1990) propuseram modelos para a estruturação de diversas seções de um texto científico, como o resumo, a introdução e a conclusão. Já Swales & Feak (1994) propuseram um modelo para a estruturação de introduções, que posteriormente foi adaptado por outros pesquisadores para outras seções (Anthony & Lashkia, 2003; Teufel & Moens, 2002). Com exceção do trabalho de Teufel & Moens (2002), esse estudos tiveram como motivação auxiliar a escrita de textos científicos, uma tarefa que é reconhecidamente difícil, especialmente para escritores iniciantes.

Do ponto de vista computacional, os estudos que tratam da classificação retórica de textos buscam construir ferramentas capazes de identificar componentes retóricos de forma automática, tendo como base um modelo de estrutura retórica que pode se aplicar ao texto completo ou apenas a uma de suas seções. Dada a importância dos resumos (abstract) para a indexação de artigos científicos em bases de dados especializadas, bem como para a realização de mapeamentos sistemáticos, várias pesquisas focam a classificação retórica de resumos (Anthony & Lashkia, 2003; Hirohata et al., 2008; Guo et al., 2011; Dayrell et al., 2012; Yepes et al., 2013; Moura, 2018). Assim como noutras áreas do PLN, a maioria dos trabalhos focam a língua inglesa. Para a língua portuguesa, destacam-se os trabalhos de Feltrim et al. (2004) e Andreani & Feltrim (2015), ambos relacionados ao classificador AZPort.

Nesse contexto, o estudo aqui apresentado teve por objetivo avaliar diferentes conjuntos de atributos e algoritmos de classificação aplicados à construção de classificadores retóricos sentenciais para resumos científicos escritos em português. Os *corpora* usados no desenvolvimento foram os mesmos usados por Feltrim et al. (2004) e Andreani & Feltrim (2015), ambos compostos de resumos de teses e dissertações na área de Ciência da Computação, manualmente anotados de acordo com um modelo de sete categorias retóricas.

Na Figura 1, é apresentado um resumo anotado extraído do *corpus* de Feltrim et al. (2004). Como é possível observar, a unidade mínima de anotação é uma sentença e nem todas as categorias retóricas possíveis aparecem no resumo. De fato, embora o modelo retórico preveja sete categorias em uma ordem específica, os resumos não apresentam obrigatoriamente todas as categorias e nem seguem uma ordem estrita entre elas.

Os atributos avaliados no estudo incluíram valores TF-IDF, *word embeddings* e os atributos usados pelo classificador AZPort. Foram avaliados diferentes modelos de *word embeddings*, com variações do número de dimensões e da estratégia de geração dos *embeddings* das sentenças. Os classificadores foram induzidos por diferentes algoritmos supervisionados e com diferentes configurações de atributos. A análise dos resultados mostrou que os atributos usados pelo AZPort foram fundamentais para o desempenho dos classificadores, mas que a combinação com outros atributos, em particular com *word embeddings*, foi benéfica.

O restante deste artigo está organizado como segue. Na Seção 2 é apresentada uma visão geral da área, bem como a descrição dos trabalhos relacionados à classificação retórica de resumos científicos em português. Em seguida, na Seção 3.1 são descritos os *corpora* usados neste estudo, bem como os atributos extraídos e os classificadores avaliados. Os resultados experimentais são apresentados e analisados na Seção 4. O desempenho dos classificadores foi analisado focando nos seguintes aspectos: combinação de atributos, modelos de *word embeddings*, dimensão do vetor de *word embeddings* e estratégia para obtenção do *embedding* de uma sentença. Por fim, na Seção 5 são feitas as conclusões a respeito do estudo e apresentadas direções para trabalhos futuros.

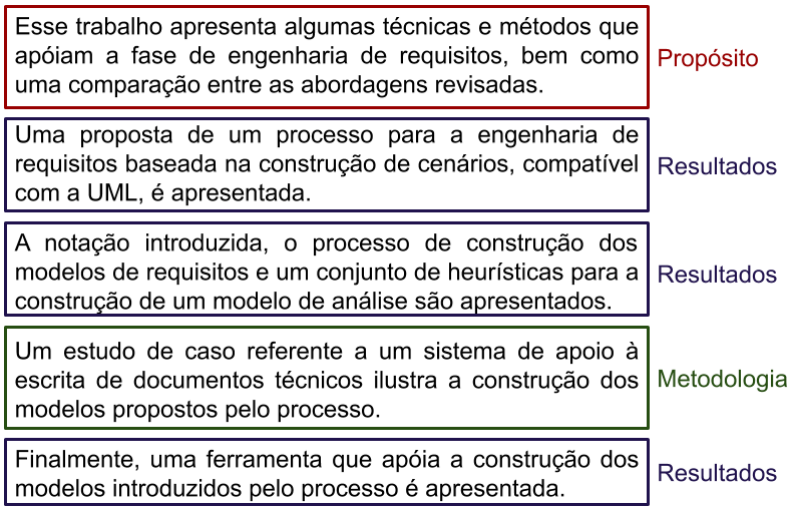

Figura 1: Exemplo de classificação de estrutura retórica das sentenças de um resumo.

2 Trabalhos relacionados

Na literatura, a classificação de estrutura retórica em textos científicos tem sido tratada como um problema de classificação sentencial, no qual se busca associar categorias retóricas às sentenças de um texto. Embora tal tarefa tenha sido abordada como um problema de classificação multirrótulo por Dayrell et al. (2012), a maioria dos trabalhos relacionados busca associar uma única categoria a cada sentença.

Com relação aos métodos de aprendizado, destaca-se o uso de métodos supervisionados e, portanto, dependentes de *corpora* anotados. Exceções são os trabalhos de Guo et al. (2011) e Guo et al. (2013), nos quais foram usados métodos semissupervisionados, e Varga et al. (2012) e Reichart & Korhonen (2012), os quais propuseram soluções não-supervisionadas.

Há uma grande variação com relação aos atributos usados para a classificação, sendo eles, em sua maioria, baseados em informações superficiais de estrutura, léxicas e morfossintáticas (Anthony & Lashkia, 2003; Mullen et al., 2005; Hirohata et al., 2008; Pendar & Cotos, 2008; Merity et al., 2009; Guo et al., 2011; Liakata et al., 2012; Yepes et al., 2013; Fisas et al., 2015). Apenas Teufel & Moens (2002) propuseram o uso de informação semântica, que se deu por meio da identificação de padrões referentes a agentes e ações.

Com relação ao modelo que define as categorias retóricas a serem identificadas pelos classificadores, também há uma grande variação entre os trabalhos encontrados, pois tais modelos costumam ser ajustados ao contexto das aplicações pretendidas por cada pesquisa. Enquanto alguns modelam os movimentos retóricos do texto como um todo (Teufel & Moens, 2002; Merity et al., 2009; Liakata et al., 2012; Varga et al., 2012; Fisas et al., 2015), outros modelam uma seção específica, como o resumo (Anthony & Lashkia, 2003; Hirohata et al., 2008; Guo et al., 2011; Dayrell et al., 2012; Reichart & Korhonen, 2012; Yepes et al., 2013) e a introdução (Pendar & Cotos, 2008).

A seguir são apresentados os trabalhos de Feltrim et al. (2004) e Andreani & Feltrim (2015), pois ambos abordaram a classificação da estrutura retórica de resumos em português. Também é apresentado o trabalho de Teufel & Moens (2002), uma vez que o mesmo serviu de ponto de partida para diversos trabalhos na área, incluindo o de Feltrim et al. (2004).

Teufel & Moens (2002) propuseram a segmentação de um artigo científico em zonas argumentativas, que juntas compõem a sua estrutura retórica. Esse modelo foi chamado de *Argumentative Zoning* (AZ). A motivação para a criação do AZ foi a sua aplicação na sumarização automática de artigos científicos. As zonas, ou categorias, previstas pelo AZ são as seguintes: Objetivo (objetivo específico do artigo), Textual (descrição da estrutura da seção), Próprio (descrição neutra de metodologia, resultados e discussão do artigo), Contexto (conhecimento científico aceito), Contraste (comparações ou diferenças com outros trabalhos, explicitando seus pontos fracos), Base (conformidades com ou continuações de outros trabalhos) e Outro (descrição neutra de trabalhos de outros pesquisadores).

Para a classificação automática, o AZ utilizou um classificador *Naive Bayes* e os atributos extraídos incluíram informações superficiais, gramaticais e semânticas. O *corpus* utilizado para aprendizado foi composto por 80 artigos,

totalizando 12.188 sentenças, e o classificador foi avaliado por meio de validação cruzada de 10 partições. Os resultados gerais obtidos pelo classificador AZ foram 50% de Macro-F, 0,45 de *Kappa* e 73% de acurácia.

Tendo por base o trabalho de Teufel & Moens (2002), Feltrim et al. (2004) propuseram o AZPort (*Argumentative Zoning for Portuguese*), uma adaptação do AZ para a língua portuguesa com foco na classificação retórica de resumos científicos. Uma vez que o AZPort teve como motivação a sua utilização em uma ferramenta de auxílio à escrita, o modelo retórico usado pelo classificador foi adaptado para esse contexto, resultando no seguinte conjunto de categorias: Contexto (B), Lacuna (G), Propósito (P), Método (M), Resultado (R), Conclusão (C) e Estrutura (O).

O AZPort utiliza um conjunto de oito atributos derivados dos atributos propostos por Teufel & Moens (2002), conforme descritos na Tabela 1. Assim como o AZ, o AZPort é um classificador bayesiano, de modo que foi necessário utilizar um conjunto de exemplos para o treinamento do modelo. O *corpus* utilizado para esse fim foi chamado de CorpusDT, sendo composto por 52 resumos que totalizam 366 sentenças, conforme descrito na Seção 3.1. O AZPort foi avaliado por meio de 13 rodadas de validação cruzada de 13 partições, obtendo 60% de Macro-F, 0,65 de *Kappa* e 72% de acurácia.

Andreani (2017) realizou um estudo acerca da aplicação de algoritmos de predição estruturada à tarefa em questão. A fim de encontrar o melhor algoritmo para a classificação de estrutura retórica, os seguintes algoritmos foram avaliados: Modelo Oculto de Markov (HMM), Modelo de Markov de Entropia Máxima (MEMM), *Conditional Random Fields* (CRF) e *Structured Support Vector Machines* (SSVM). Para que fosse possível realizar a comparação com o AZPort, foi usado o mesmo modelo de estrutura retórica (B, G, P, M, R, C e O) e dois *corpora* foram empregados no treinamento e teste dos classificadores: o CorpusDT e o *Corpus* 466 (Andreani & Feltrim, 2015). Assim como o CorpusDT, o *Corpus* 466 é composto por resumos extraídos de teses e dissertações em Ciência da Computação, totalizando 466 sentenças manualmente anotadas.

Os resultados de Andreani (2017) foram calculados a partir de 30 execuções de validação cruzada de 13 partições. Os atributos utilizados incluíram os atributos do AZPort exceto Citação e Histórico, *n-gramas*, segmentação e janela deslizante. O atributo *n-gramas* é composto por valores TF-IDF; o atributo segmentação indica o início, meio e fim de segmentos (sequências de sentenças) com uma mesma categoria retórica; e, por fim, a janela deslizante inclui os atributos das $k = \{0, 1, 2\}$ sentenças vizinhas. Os resultados experimentais mostraram que o melhor desempenho foi obtido pelo classificador CRF, com F1-*score* de 68%, o que representou uma melhoria de 7% em relação ao desempenho do AZPort.

3 Desenvolvimento

Nesta seção são descritos os *corpora* empregados neste estudo, bem como os atributos e classificadores avaliados.

3.1 *Corpora*

Conforme mencionado anteriormente, os *corpora* utilizados neste estudo são constituídos de resumos de teses e dissertações da área de Ciência da Computação, escritos em português do Brasil. Esses resumos foram coletados e anotados como parte dos trabalhos de Feltrim et al. (2004) e Andreani & Feltrim (2015). Todas as sentenças foram anotadas manualmente por três anotadores treinados e com experiência em escrita científica. A concordância entre os anotadores medida por meio da estatística *Kappa* (Siegel & Castellan Jr., 1988) foi de 0,695.

Os *corpora*, que neste estudo foram chamados de 366, 466 e 832, estão anotados com as categorias previstas pelo classificador AZPort, sendo elas: Contexto, Lacuna, Propósito, Resultado, Método, Conclusão e Estrutura. Os *corpora* 366 (chamado por Feltrim et al. (2004) de CorpusDT) e 466 possuem 52 resumos cada, totalizando, respectivamente, 366 e 466 sentenças. O *corpus* 832 corresponde a união dos *corpora* 366 e 466 e, consequentemente, possui 104 resumos, totalizando 832 sentenças.

A distribuição de categorias observada nos *corpora* é mostrada na Tabela 2. Como se pode notar, a distribuição é semelhante nos *corpora* 366 e 466, sendo as categorias Contexto e Resultado as mais frequentes e as categorias Conclusão e Estrutura as menos frequentes. A prevalência da categoria Resultado é comum em *corpora* de resumos científicos, uma vez que os autores buscam enfatizar os resultados encontrados em seus trabalhos (Hirohata et al., 2008; Moura, 2018). Já a alta frequência da categoria Contexto é uma característica particular dos *corpora* usados neste trabalho e se deve ao fato de os resumos terem sido extraídos a partir de teses e dissertações. Os resumos desses tipos de trabalhos tendem a ser mais longos, permitindo que os autores contex-

Atributo	Descrição	Valores possíveis
Tamanho	Qual é o tamanho da sentença em comparação aos limiares 20 e 40 palavras?	curta, média ou longa
Localização	Qual é a posição da sentença no resumo?	primeira, segunda, mediana, penúltima ou última
Citação	A sentença contém citações?	sim ou não
Expressão	Que tipo de expressão padrão a sentença contém?	B, C, G, M, P, R ou *noExpr*
Tempo	Qual é o tempo do primeiro verbo finito da sentença?	IMP, PRES, PAST, FUT, PRES-CPO, PAST-CPO, FUT-CPO, PRES-CT, PAST-CT, FUT-CT, PRES-CPO-CT, PAST-CPO-CT, FUT-CPO-CT ou *noVerb*
Voz	Qual é a voz do primeiro verbo finito da sentença?	passiva, ativa ou *noVerb*
Modal	O primeiro verbo finito da sentença é modal?	sim, não ou *noVerb*
Histórico	Qual é a categoria da sentença anterior?	_, B, C, G, M, O, P ou R

Tabela 1: Conjunto de atributos utilizado do AZPort (adaptado de Feltrim (2004)).

tualizem melhor suas áreas de pesquisa. Com relação a categoria Estrutura, que é minoritária, cabe destacar que é incomum que informações a respeito da organização do texto sejam incluídas em resumos, o que justifica a baixa frequência de sentenças dessa categoria nos *corpora*.

3.2 Atributos

Foram utlizados todos os atributos usados pelo AZPort (Tabela 1) com a adição de um novo atributo que registra a posição relativa da sentença no resumo. Também foram extraídos atributos por meio de TF-IDF e *word embeddings*.

A extração dos vetores com valores TF-IDF (*Term Frequency–Inverse Document Frequency*) foi feita com base em unigramas. Em seguida, os 100 melhores atributos foram selecionados por meio do teste χ^2 (qui-quadrado). A escolha pelo uso dos 100 melhores unigramas se deu por experimentação. Foram avaliadas diferentes configurações de vetores resultantes das combinações de diferentes valores de n (1, 2 e 3) para os n-gramas e diferentes valores de corte (50, 100, 250, 500 e 1000) para o teste χ^2.

Os atributos baseados em *word embeddings* (WE) foram extraídos utilizando-se os modelos CBOW, *Skip-gram* — ambos com as ferramentas Word2Vec (Mikolov et al., 2013) e Wang2Vec (Ling et al., 2015) — e GloVe (Pennington et al., 2014), todos eles previamente treinados para o português por Hartmann et al. (2017). Tais modelos estão disponíveis no repositório NILC–*Embeddings*[1] do NILC (ICMC–USP)[2] e foram gerados a partir de um *corpus* em português brasileiro e europeu, contendo textos de fontes e gêneros variados (Hartmann et al., 2017). Cada modelo está disponível no repositório com vetores de 50, 100, 300, 600 e 1000 dimensões.

Modelos de *word embeddings*, todavia, retornam vetores para palavras e, neste trabalho, a unidade de classificação é uma sentença. Dessa forma, os *embeddings* das sentenças foram gerados por meio de duas estratégias — média simples e média ponderada pelo valor IDF (*Inverse Document Frequency*). Assim, o *embedding* de uma sentença correspondeu à média simples ou à média ponderada dos *word embeddings* das palavras que a formavam.

Além desses atributos serem utilizados de maneira individual, também foram feitas combinações entre eles. Por meio da concatenação dos vetores extraídos de cada um dos atributos, foram feitas as seguintes combinações:

- TF-IDF + atributos AZPort;
- *Word embeddings* + TF-IDF;
- *Word embeddings* + atributos AZPort; e
- *Word embeddings* + TF-IDF + atributos AZPort.

[1] NILC–*Embeddings*. http://www.nilc.icmc.usp.br/embeddings

[2] Núcleo Interinstitucional de Linguística Computacional. http://nilc.icmc.usp.br

Categoria	***Corpus* 366**	***Corpus* 466**	***Corpus* 832**
Contexto	77	179	256
Lacuna	36	36	72
Propósito	65	68	133
Método	45	59	104
Resultado	117	103	220
Conclusão	20	20	40
Estrutura	6	1	7
Total	**366**	**466**	**832**

Tabela 2: Número de sentenças de cada *corpus* (adaptado de Andreani (2017)).

Em todas as combinações de atributos avaliadas, foram adicionados à representação de uma sentença s_i os atributos das sentenças s_{i-1} e s_{i+1} sempre que possível — quando a sentença é a primeira do resumo, não há uma sentença s_{i-1}; e quando a sentença é a última, não há uma sentença s_{i+1}.

3.3 Classificadores

Os seguintes classificadores foram usados: *k-nearest neighbors* (K-NN), *Naive Bayes* — com as variações Gaussiana (G-NB) e Bernoulli (B-NB) —, *Decision Trees* (DT) — com a implementação do algoritmo CART—, *Support Vector Machines* (SVM) — com *kernels* linear e *Radial-Basis Function* (RBF) — e *Conditional Random Fields* (CRF). Com exceção do algoritmo CRF, foram usadas as implementações fornecidas pelas bibliotecas *scikit-learn*[3]. Para o CRF foi utilizada a biblioteca *sklearn-crfsuite*[4], uma versão da ferramenta *CRFsuite*[5] que é compatível com os estimadores da *scikit-learn*.

Para o algoritmo K-NN, foram considerados três vizinhos ($n\_neighbors = 3$) e distância Euclidiana; o restante dos parâmetros foram usados com valores *default*. Os algoritmos bayesianos, G-NB e B-NB, foram utilizados com seus parâmetros *default*. No DT, o único parâmetro com valor alterado foi $random\_state = 0$, o qual determina a semente para gerar números aleatórios.

Para o classificador SVM, foram alterados os seguintes parâmetros: *kernel*, C (parâmetro de penalidade do termo de erro) e *gamma* (coeficiente de *kernel*, no caso, apenas para RBF). O primeiro, *kernel* =‘*linear*’ ou *kernel* =‘*rbf*’, determina se o *kernel* a ser usado é linear ou RBF, respectivamente. No SVM-linear, para os *corpora* 366 e 832, foi usado $C = 100$; e, para o *corpus* 466, $C = 1000$. Já no SVM-RBF, para todos os *corpora*, foram usados $C = 1000$ e $gamma = 0.001$. Esses parâmetros foram escolhidos por meio da execução do algoritmo *Grid Search* para maximizar o desempenho com base nos atributos TF-IDF.

Para o classificador CRF, os parâmetros usados foram os seguintes: *algorithm = ‘lbfgs’, c1 = 0.1, c2 = 0.1, max_iterations = 100, all_possible_transitions = True*. O parâmetro *algorithm* especifica o algoritmo de treinamento, neste caso, gradiente descente usando o método L-BFGS (*default*); os parâmetros *c1* e *c2* definem os coeficientes de regularização L1 e L2, respectivamente; *max_iterations* define o número máximo de iterações para a otimização; e *all_possible_transitions = True* especifica que todas as transições possíveis devem ser geradas, mesmo as que não ocorrem no conjunto de treinamento. O restante dos parâmetros foram usados com seus valores *default*.

4 Resultados

Nesta seção são apresentados os resultados dos experimentos realizados com os *corpora* e os classificadores descritos. Em todos os experimentos, os classificadores foram avaliados por meio de validação cruzada estratificada de 10 partições. Embora a classificação seja feita por sentença, a geração das partições foi feita a partir de um conjunto de resumos, de modo a garantir que todas as sentenças de um mesmo resumo estejam em uma mesma partição. Vale destacar ainda que a mesma divisão de partições foi usada na avaliação de todos os classificadores.

A Tabela 3 apresenta os melhores valores de medida F1 obtidos por cada classificador usando cada uma das combinações de atributos. Nessa tabela e nos gráficos desta seção, RBF corresponde ao classificador SVM com *kernel* RBF e SVM corresponde ao classificador SVM com *ker-*

[3] http://scikit-learn.org

[4] https://sklearn-crfsuite.readthedocs.io/en/latest/

[5] http://www.chokkan.org/software/crfsuite/

Atributos	**RBF**	**SVM**	**K-NN**	**G-NB**	**B-NB**	**DT**	**CRF**
TF-IDF	46%	57%	39%	23%	45%	46%	56%
WE	50%	49%	37%	**74%**	44%	36%	55%
AZPort	66%	66%	56%	26%	64%	**62%**	92%
TF-IDF + AZPort	66%	62%	51%	33%	59%	61%	92%
WE + TF-IDF	54%	54%	38%	33%	49%	43%	58%
WE + AZPort	**71%**	**71%**	**61%**	59%	**68%**	60%	**94%**
Todos os atributos	**71%**	69%	53%	41%	62%	60%	**94%**
Média	61%	61%	48%	41%	56%	53%	77%
Desvio Padrão	10%	8%	10%	19%	10%	11%	20%

Tabela 3: Melhores resultados obtidos por classificador e combinação de atributos.

nel linear. O melhor desempenho obtido por cada algoritmo está destacado em negrito.

O melhor desempenho do SVM-RBF (71%) foi obtido no *corpus* 466 com as seguintes combinações: WE (Wang2Vec–*Skip-gram*–600) com atributos do AZPort e todos os atributos (GloVe–1000 combinado com os atributos do AZPort e TF-IDF). Na primeira combinação, os *embeddings* das sentenças foram calculados por média ponderada e, na segunda, por média simples. O SVM-linear também teve seu melhor desempenho (71%) obtido no *corpus* 466, porém apenas para a combinação de WE com atributos do AZPort. Nesse caso, o modelo de WE usado foi Wang2Vec–*Skip-gram*–600 e os *embeddings* das sentenças foram obtidos pela média simples.

O K-NN obteve o menor dos melhores desempenhos, 63%. Esse resultado foi obtido no *corpus* 466 com a combinação de WE com os atributos do AZPort. Os modelos de WE nesse caso foram dois: Wang2Vec–Skip-gram de dimensão 600 com a média ponderada e GloVe de dimensões 50 e 100 com a média simples.

Os algoritmos bayesianos alcançaram 74% e 68% com as variações gaussiana (G-NB) e Bernoulli (B-NB), respectivamente. Vale notar que o G-NB atingiu o segundo melhor desempenho entre os algoritmos avaliados utilizando apenas os atributos WE. Esse resultado foi alcançado com o modelo Word2Vec–CBOW–1000 com *embeddings* gerados pelas médias ponderada e simples no *corpus* 366 e com o mesmo modelo, mas apenas com a média ponderada, para o *corpus* 466. Já o B-NB atingiu 68% para o *corpus* 466, com os modelos Word2Vec–*Skip-gram*–50 e GloVe–50 usando médias simples e ponderada, respectivamente.

O melhor desempenho do DT (62%) também foi obtido no *corpus* 466, porém utilizando apenas os atributos do AZPort. Esse foi o único classificador cujo melhor resultado que não incluiu WE.

O melhor desempenho observado (94%) foi obtido pelo classificador CRF no *corpus* 466. Esse resultado foi alcançado usando a combinação de WE (Wang2Vec–*Skip-gram*–100 com média simples e Wang2Vec–*Skip-gram* e GloVe, ambos de dimensão 300, com média ponderada) com os atributos do AZPort. O mesmo desempenho foi obtido usando-se todos os atributos (Word2Vec–*Skip-gram* de dimensões 100 e 300 com média ponderada; e GloVe–300 com médias ponderada e simples).

A Tabela 4 mostra os valores de precisão, revocação e F1 obtidos pelo classificador CRF com a combinação WE (Wang2Vec–*Skip-gram*–300 com média ponderada) e AZPort para o *corpus* 466. É possível observar que o classificador mantém o desempenho acima de 85% para todas as categorias exceto Conclusão e Estrutura. No caso da categoria Estrutura, o desempenho foi nulo devido à ausência de sentenças dessa categoria, uma vez que há apenas uma sentença com essa classificação no *corpus* 466. Já no caso da categoria Conclusão, além da sua baixa frequência no *corpus*, existe uma dificuldade por parte do classificador em distinguir as categorias Conclusão e Resultado.

Categoria	**Precisão**	**Revocação**	**F1**
Contexto	99%	100%	100%
Lacuna	100%	100%	100%
Propósito	96%	94%	95%
Método	93%	92%	92%
Resultado	85%	91%	88%
Conclusão	67%	50%	57%
Estrutura	0%	0%	0%
Média	93%	94%	94%

Tabela 4: Precisão, revocação e F1 obtidas pelo classificador CRF usando WE + AZPort sobre o *corpus* 466.

Nas últimas duas linhas da Tabela 3 são apresentadas as médias e desvios padrões dos desempenhos dos algoritmos. Observa-se que o melhor desempenho médio (77%) foi obtido pelo CRF, que também foi o classificador com maior dispersão (desvio padrão igual a 20%). Os algoritmos SVM (RBF e linear) tiveram desempenhos médios iguais (61%), ambos com valores baixos de desvio padrão (10% e 8%, respectivamente).

Ainda na Tabela 3, também é possível observar que os atributos TF-IDF, usados de forma isolada, não trouxeram bons resultados. Além disso, esses atributos não mostraram ter influência no desempenho de algoritmos como SVM-RBF e CRF. Esses dois algoritmos mantiveram o desempenho independentemente da adição dos valores TF-IDF no conjunto de atributos. Ainda, para os classificadores SVM-Linear e B-NB, o desempenho piorou com a adição dos atributos TF-IDF.

A utilização dos atributos do AZPort levaram às melhorias mais significativas em termos dos desempenhos dos classificadores. No caso do CRF, os atributos do AZPort se mostraram fundamentais, já que utilizando apenas esses atributos o CRF atingiu 92% de F1 para o *corpus* 466. Tal valor ficou apenas 2% abaixo do melhores desempenhos observados.

Atributos	**Média**	**Desv.Padrão**
TF-IDF	45%	11%
WE	49%	13%
AZPort	62%	19%
TF-IDF + AZPort	61%	18%
WE + TF-IDF	47%	9%
WE + AZPort	69%	12%
Todos os atributos	64%	17%

Tabela 5: Média e Desvio Padrão dos melhores desempenhos obtidos com cada combinação de atributos.

A Tabela 5 mostra a média e o desvio padrão dos melhores desempenhos obtidos com cada combinação de atributos. Novamente é evidenciada a contribuição dos atributos do AZPort com média superior aos dos outros atributos usados de forma isolada. Quando incluídos no conjunto de atributos, as médias são superiores a 60%, enquanto que, na sua ausência, as médias não alcançam 50%. Vale notar ainda que, embora não tenham tido um bom desempenho médio quando usados de forma isolada, os atributos WE melhoraram o desempenho da classificação quando adicionados aos atributos do AZPort. A maior média de desempenho, 69%, foi obtida por essa combinação.

Para uma melhor visualização dos resultados obtidos por *corpora*, a Figura 2 mostra um gráfico com os maiores valores de F1 obtidos por cada classificador em cada um dos *corpora*. Contrariando o esperado, os resultados não foram melhores no maior *corpus*. No geral, o que gerou os melhores resultados em todos os classificadores foi o *corpus* 466. Já o *corpus* 366 se mostrou o mais difícil para os classificadores. Apenas o G-NB obteve os melhores resultados no *corpus* 366, junto ao 466. Uma análise mais aprofundada dos *corpora* é necessária para identificar as razões que levaram a esses resultados, mas a dificuldade evidenciada para o *corpus* 366 possivelmente influenciou os resultados obtidos para o *corpus* 832, fazendo com que os mesmos ficassem abaixo dos obtidos para o *corpus* 466 apesar do maior número de sentenças.

Com relação aos modelos de *word embeddings*, o gráfico da Figura 3 mostra as maiores porcentagens obtidas com os cinco modelos avaliados (Word2Vec–CBOW, Word2Vec–*Skip-gram*, Wang2Vec–CBOW, Wang2Vec–*Skip-gram* e GloVe). Nesse gráfico, as porcentagens correspondem às maiores medidas F1 obtidas usando-se apenas WE como atributos.

Na Figura 3 é possível notar que o modelo Wang2Vec–*Skip-gram* foi o melhor modelo apenas para o K-NN, junto com Word2Vec–*Skip-gram* e o outro modelo Wang2Vec. Além disso, foi o pior para os classificadores SVM (RBF e linear) e DT e foi o segundo pior para o CRF. Curiosamente, quatro das combinações nas quais o CRF atingiu seu melhor desempenho, 94%, utilizou o modelo Wang2Vec–*Skip-gram*. Outro ponto a ser destacado é que nenhum classificador apresentou melhor desempenho com o modelo GloVe, mas as outras três combinações em que o CRF obteve 94% incluem os atributos extraídos a partir do modelo GloVe.

O gráfico da Figura 4 mostra um comparativo de desempenho para os diferentes tamanhos dimensionais de WE avaliados. Da mesma forma, as porcentagens correspondem às maiores medidas F1 obtidas usando-se apenas WE como atributos. Para o algoritmo K-NN, a variação da dimensão dos vetores de *word embeddings* não causou variações significantes nos resultados. Já para o DT, o aumento da dimensão piorou o desempenho. Para o restante dos classificadores, pode-se dizer que o aumento da dimensão mostrou melhora no desempenho. O G-NB foi o classificador que mostrou ter maior proporção de melhora em comparação aos outros.

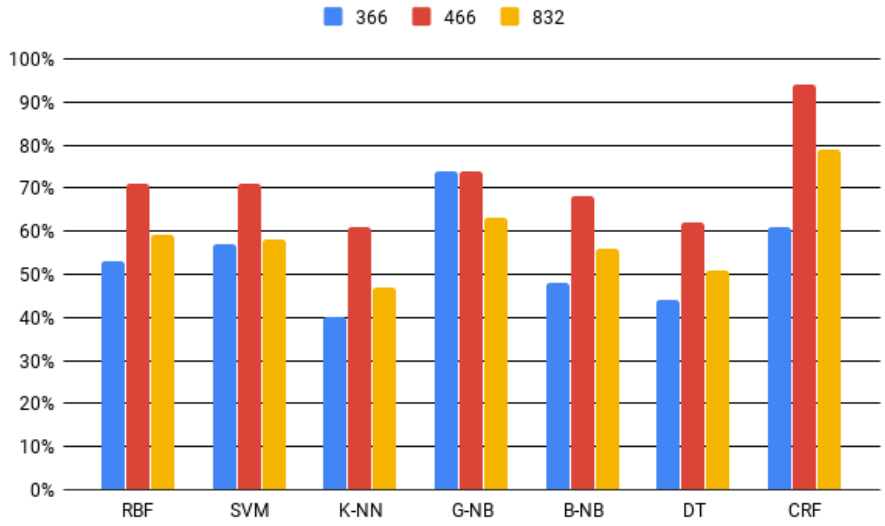

Figura 2: Melhores resultados obtidos para cada *corpus*.

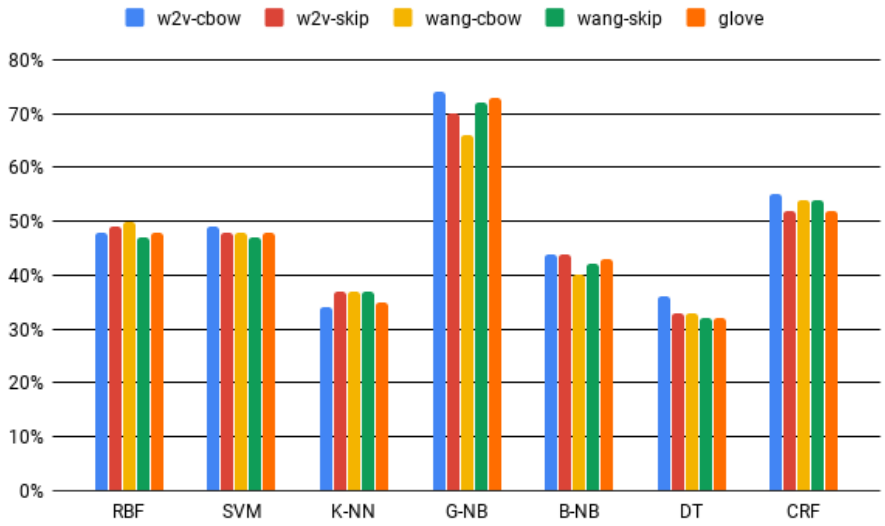

Figura 3: Melhores resultados obtidos em cada modelo de *word embedding*.

O gráfico da Figura 5 mostra um comparativo dos melhores resultados obtidos com as duas formas de representação de *embeddings* para uma sentença (média ponderada pelo IDF e média simples). Nesse gráfico, as porcentagens também correspondem às maiores medidas F1 obtidas usando-se apenas WE como atributos. É possível observar que houve pouca variação no desempenho dos classificadores devido à representação usada. O classificador G-NB se mostrou indiferente quanto à forma de representação, enquanto o restante dos classificadores mostraram resultados levemente superiores usando a média simples. Cabe destacar, no entanto, que quatro das sete combinações em que o CRF obteve 94% usou *embeddings* gerados pela média ponderada.

5 Conclusão

Neste estudo foram avaliados diferentes conjuntos de atributos e algoritmos de classificação aplicados à construção de classificadores retóricos sentenciais para resumos científicos escritos em português. Foram avaliados atributos baseados em TF-IDF, *word embeddings* e os atributos do classificador AZPort, bem como as suas combinações. Com relação aos *embeddings*, foram avaliados os diferentes modelos disponibilizados no repositório NILC–*Embeddings*, bem como duas estratégias de geração de *embeddings* de sentenças a partir de *word embeddings*. As diferentes configurações de atributos foram avaliadas em combinação com sete classificadores distintos, todos eles supervisionados.

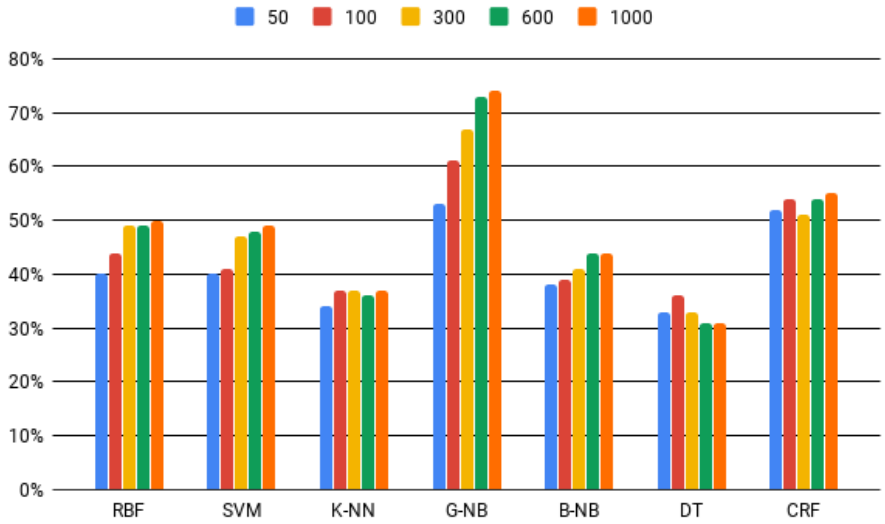

Figura 4: Melhores resultados obtidos em cada dimensão do vetor de *word embedding*.

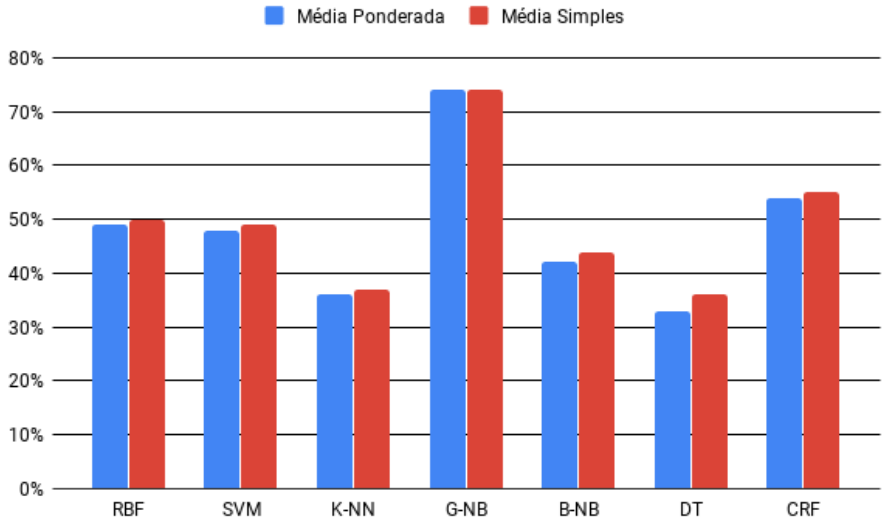

Figura 5: Melhores resultados obtidos em cada forma de representação de *word embedding* para uma sentença.

Dentre os algoritmos avaliados, o que obteve melhor desempenho foi o CRF, confirmando que a classificação retórica é uma tarefa apropriada para algoritmos de rotulação sequencial (Andreani & Feltrim, 2015). No entanto, os resultados se mostraram mais dependentes do conjunto de atributos utilizado.

Observou-se que a utilização de WE, individualmente, não trouxe ganhos significativos nos desempenhos dos classificadores em relação à utilização individual de TF-IDF, com exceção do classificador G-NB, que atingiu o seu melhor desempenho com esses atributos. Já a utilização individual dos atributos do AZPort mostrou desempenho superior ao desempenho com a utilização de TF-IDF e WE em mais de 20% para o CRF e entre 9% a 19% para os classificadores SVM (RBF e linear), K-NN, B-NB e DT. Em especial, o classificador DT obteve seus melhores resultados usando apenas os atributos do AZPort. No caso do classificador CRF, o atributos do AZPort se mostraram os mais efetivos, já que, apenas com eles, o CRF alcançou 92% de F1 para um dos *corpora* usados. Isso mostra que atributos que codificam informações além da superfície do texto, como os do AZPort, são importantes para classificação de estrutura retórica, especialmente quando o conjunto de treinamento é reduzido, como foi o caso deste trabalho.

A combinação de WE com outros atributos mostrou utilidade, uma vez que os classificadores SVM-linear, K-NN e B-NB obtiveram seus melhores resultados com a combinação WE com os atributos do AZPort. Outros classificadores, como SVM-RBF e CRF, obtiveram seus melhores resultados tanto com a combinação de WE com os atributos do AZPort quanto com a combinações de todos os atributos.

Considerando-se os desempenhos obtidos pelos modelos de *word embeddings* quando usados de forma individual, o modelo que obteve o melhor resultado médio foi o Word2Vec–CBOW. Esse resultado está de acordo com o trabalho de Sousa (2016), que também destacou o modelo CBOW como tendo melhor desempenho. Entretanto, o melhor resultado de classificação usando o CRF foi obtido com os modelos Wang2Vec–*Skip-gram* e GloVe. Esse resultado está de acordo com o trabalho de Hartmann et al. (2017) que destacou o desempenho do modelo Wang2Vec. O aumento das dimensões dos modelos de WE trouxe melhorias ao desempenhos dos classificadores, com exceção do K-NN. Essa melhoria ocorreu em maior proporção para o classificador G-NB do que para os outros classificadores.

Para a representação do *embedding* de uma sentença, embora seja comum na literatura realizar a combinação dos *word embeddings* por meio da média ponderada pelo IDF, para este estudo a melhor estratégia foi a combinação pela média simples. Conforme mostrado na Figura 5, a maioria dos classificadores obtiveram melhores resultados com tal estratégia.

Uma vez que os atributos gerados a partir de *word embeddings* contribuíram para a melhoria dos resultados, ainda que de forma discreta, uma das direções para trabalhos futuros é o estudo de outras formas de representação para os *embeddings* das sentenças. Outro ponto para investigações futuras é o treinamento de modelos de *embeddings* com *corpus* de domínio científico, uma vez que os modelos avaliados neste estudo foram treinados com *corpora* de domínios variados. Cabe destacar que, apesar da diferença de domínio, a taxa de palavras não encontradas nos modelos de *word embeddings* variou entre 11% e 13% para os três *corpora* usados no estudo.

Ainda com relação aos atributos, outra vertente de trabalhos futuros é a proposta e a avaliação de atributos que codifiquem informações semânticas, como as fornecidas por analisadores semântico (*Semantic Role Labeling* – SRL).

Agradecimentos

As autoras agradecem aos revisores pelas importantes contribuições a este estudo.

Referências

Andreani, Alexandre C. 2017. Predição estruturada aplicada à detecção de estrutura retórica. Dissertação (Pós Graduação em Ciência da Computação), Universidade Estadual do Paraná, Maringá, Brazil.

Andreani, Alexandre C. & Valéria D. Feltrim. 2015. Campos aleatórios condicionais aplicados à detecção de estrutura retórica em resumos de textos acadêmicos em português (conditional random fields applied to rhetorical structure detection in academic abstracts in portuguese). Em *10th Brazilian Symposium in Information and Human Language Technology (STIL)*, 111–120.

Anthony, Laurence & George V. Lashkia. 2003. Mover: a machine learning tool to assist in the reading and writing of technical papers. *IEEE Transactions on Professional Communication* 46(3). 185–193. 10.1109/TPC.2003.816789.

Booth, Wayne C., Gregory G. Colomb, Joseph M. Wiliams & Henrique A. Rego Monteiro. 2005. *A arte da pesquisa.* São Paulo: Martins Fontes.

Dayrell, Carmen, Arnaldo Candido Jr., Gabriel Lima, Danilo Machado Jr., Ann Copestake, Valéria D. Feltrim, Stella Tagnin & Sandra M. Aluísio. 2012. Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. Em *8th International Conference on Language Resources and Evaluation (LREC)*, 1604–1609.

Feltrim, Valéria D. 2004. *Uma abordagem baseada em corpus e em sistemas de crítica para a construção de ambientes web de auxilio à escrita acadêmica em português*: Universidade de São Paulo, São Carlos, Brazil. Tese de Doutoramento.

Feltrim, Valéria D., Jorge M. Pelizzoni, Simone Teufel, Maria das Graças Volpe das Nunes & Sandra M. Aluísio. 2004. Applying argumentative zoning in an automatic critiquer of academic writing. Em *Advances in Artificial Intelligence – SBIA 2004*, vol. 3171, 214–223. Springer Berlin Heidelberg. 10.1007/978-3-540-28645-5_22.

Fisas, Beatriz, Francesco Ronzano & Horacio Saggion. 2015. On the discoursive structure of computer graphics research papers. Em *9th Linguistic Annotation Workshop (held in conjuncion with NAACL)*, 42–51. 10.3115/v1/W15-1605.

Guo, Yufan, Anna Korhonen & Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 273–283.

Guo, Yufan, Ilona Silins, Ulla Stenius & Anna Korhonen. 2013. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics* 29(11). 1440–1447. 10.1093/bioinformatics/btt163.

Hartmann, Nathan, Erick R. Fonseca, Christopher Shulby, Marcos Vinícius Treviso, Jessica Rodrigues & Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *CoRR* arxiv:abs/1708.06025.

Hirohata, Kenji, Naoaki Okazaki, Sophia Ananiadou & Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. Em *3rd International Joint Conference on Natural Language Processing*, 381–388.

Liakata, Maria, Shyamasree Saha, Simon Dobnik, Colin Batchelor & Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28(7). 991–1000. 10.1093/bioinformatics/bts071.

Ling, Wang, Chris Dyer, Alan W. Black & Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299–1304. 10.3115/v1/N15-1142.

Mann, William C. & Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Relatório Técnico. ISI/RS-87-190 Information Sciences Institute. 10.1515/text.1.1988.8.3.243.

Mann, William C. & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3). 243–281. 10.1515/text.1.1988.8.3.243.

Merity, Stephen, Tara Murphy & James R. Curran. 2009. Accurate argumentative zoning with maximum entropy models. Em *Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, 19–26.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* arxiv:abs/1301.3781.

Moura, Gustavo Bennemann. 2018. Redes neurais recorrentes para a classificação de estruturas retóricas. Dissertação (Pós Graduação em Ciência da Computação), Universidade Estadual do Paraná, Maringá, Brazil.

Mullen, Tony, Yoko Mizuta & Nigel Collier. 2005. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *ACM SIGKDD Explorations Newsletter* 7(1). 52–58. 10.1145/1089815.1089823.

Pendar, Nick & Elena Cotos. 2008. Automatic identification of discourse moves in scientific article introductions. Em *3rd Workshop on Innovative use of NLP for Building Educational Applications*, 62–70.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162.

Reichart, Roi & Anna Korhonen. 2012. Document and corpus level inference for unsupervised and transductive learning of information structure of scientific documents. Em *24th International Conference on Computational Linguistics (COLING)*, 995–1006.

Romeiro, Ana Karoline Queiroz. 2016. *Um estudo sobre o uso da teoria da estrutura retórica (rst) para sumarizar a sabedoria da coletividade*: Universidade Federal Fluminense. Tese de Mestrado.

Siegel, Sidney & N. John Castellan Jr. 1988. *Nonparametric statistics for the behavioral sciences*. Berkeley, CA: McGraw-Hill 2nd edn.

Sousa, Samanta de. 2016. Estudo de modelos de word embedding. Bacharel em Ciência da Computação, Universidade Tecnológica Federal do Paraná, Medianeira, Brazil.

Swales, John M. & Christine B. Feak. 1994. *Academic writing for graduate students: Essential tasks and skills: A course for nonnative speakers of english (English for specific purposes)*. Ann Arbor.

Teufel, Simone & Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics* 28(4). 409–445. doi 10.1162/089120102762671936.

Varga, Andrea, Daniel Preotiuc-Pietro & Fabio Ciravegna. 2012. Unsupervised document zone identification using probabilistic graphical models. Em *8th International Conference on Language Resources and Evaluation (LREC)*, 1610–1617.

Weissberg, Robert & Suzanne Buker. 1990. *Writing up research*. Prentice Hall Englewood Cliffs, NJ.

Yepes, Antonio Jimeno, James Mork & Alan Aronson. 2013. Using the argumentative structure of scientific literature to improve information access. Em *Workshop on Biomedical Natural Language Processing*, 102–110.

Propuesta recibida el octubre 2018 y aceptada para publicación el mayo 2019.

The Development and Evaluation of a Corpus-based Spanish Collocation Error Detection and Revision Suggestion Tool

Desarrollo y evaluación de una herramienta basada en corpus para identificar errores y realizar propuestas de corrección en las colocaciones del español

Hui-Chuan Lu
National Cheng Kung University
huichuanlu1@gmail.com

An Chung Cheng
University of Toledo
ACheng@utnet.utoledo.edu

Shujuan Wang
University of Toledo
shujuan.wang@rockets.utoledo.edu

Resumen

Durante las últimas tres décadas, el estudio de las colocaciones ha sido uno de los polos de atención por parte de las personas interesadas en el léxico, tanto en la lingüística teórica como en la aplicada. Nuestro equipo de investigación ha desarrollado y evaluado una herramienta basada en corpus para la asistencia al aprendizaje de las colocaciones en español. Basada en dos corpus (CEATE y CPEIC) y en una herramienta de extracción de colocaciones en español (HCE), esta herramienta de aprendizaje asistido por ordenador, SpColEDRS, facilita la identificación de errores en las colocaciones del español y, además, sugiere correcciones. Los resultados de la evaluación indican que la herramienta desarrollada en esta investigación puede ayudar eficazmente a los estudiantes, en especial, a los principiantes. Así mismo, los resultados de la encuesta de satisfacción confirman la utilidad de esta herramienta en el aprendizaje asistido de las colocaciones en español. Por último, este estudio arroja luz sobre las aplicaciones pedagógicas de los corpus elaborados y el aprendizaje de colocaciones en español con un enfoque basado en corpus y en un entorno de adquisición multilingüe.

Keywords

colocaciones, identificación de errores, sugerencias de corrección, español

Abstract

The topic of collocation has drawn attention for the past three decades in the lexical area of theoretical and applied linguistics. Our research team developed and evaluated a corpus-based assisted tool for collocation learning in Spanish. Based on the two constructed corpora (CEATE and CPEIC) and a Spanish collocation extraction tool (HCE), this computer-assisted learning tool, SpColEDRS, is easy to use to detect errors and also suggests revisions for Spanish collocations. Based on the evaluation, the research results indicated that the tool developed in this research can assist learners effectively, especially in the case of beginners. In addition, the results of the satisfaction survey provided positive confirmation of the effectiveness of this tool in assisting the learning of Spanish collocations. Finally, this study shed light on pedagogical applications of the constructed corpora and the learning of Spanish collocation with a corpus-based approach in a multilingual acquisition setting.

Keywords

collocation, error, detection, revision, suggestion, Spanish

1 Introduction

The topic of collocation has drawn special attention for the past thirty years in the lexical area of theoretical and applied linguistics. Wray (2000), Nattinger & DeCarrico (1992), Sinclair (1991), and Firth (1957) all indicated the importance of collocation for learning foreign languages. Many researchers have tried to define and describe collocations, but there is no one simple, precise definition of collocations. In corpus linguistics, "collocation" is defined as a group of words that co-occur more frequently than would be expected by chance (McKeown & Radev, 2000). For example, they comprise word combinations such as "boiling hot," that is, phrases that are more restricted than free combinations ("very hot") and less restricted than idioms ("get hot under the

DOI: 10.21814/lm.11.1.271
This work is Licensed under a Creative Commons Attribution 4.0 License

collar"). Correct uses of collocations could be an indication of a learner's knowledge of phrases or common combinations in the target language L2 learners because beginning learners often are not aware of the important role of collocation since they tend to focus on the learning of new words and grammatical points. Correct usage of collocation can be an obstacle even for advanced learners (Källkvist, 1995; Granger, 1998; Lorenz, 1999; Nesselhauf, 2003, 2005). In the field, compared with English collocation learning and teaching, there are fewer available tools intended to assist with learning Spanish collocations than there are for learning English (Weisser, 2016). Therefore, the purpose of this study is to further the application of previously constructed corpora and tools by developing a tool intended to assist with learning Spanish collocation. Based on a general detection and revision tool (System of Error Detection and Revision Suggestion, SEDRS) developed in 2013 (Lu et al., 2013), the corpora and the tool contained in this research comprise a learners' corpus CEATE (Corpus Escrito de Aprendices Taiwaneses de Español / Taiwanese Learners' Written Corpus of Spanish), a parallel trilingual corpus CPEIC (Corpus Paralelo de Español, Inglés y Chino / Parallel Corpus of Spanish, English and Chinese), and a Spanish collocation extraction tool, HCE (Herramienta de Colocaciones Españolas / Spanish Collocation Tool).

The primary tasks of this study include two areas of focus. The first one deals with the development of a corpus-based tool for assisting with learning Spanish collocation (a system of Spanish collocation error detection and revision suggestions), which can identify collocation errors and make correction suggestions. The second area is an empirical evaluation of the effectiveness of the developed Spanish collocation learning tool. The following are the research questions that guided the assessment of the functions of the assisted learning tool and system. (1) Are there any significant differences between the experimental and control groups after a pedagogical intervention using the corpus-based learning tool? (2) Are there any significant differences between beginning and intermediate learners in the post-test after using the learning tool?

2 Previous research

2.1 Collocation assisted learning tools

In order to obtain a general view of computer-assisted collocation learning tool, we evaluated eight existing tools used to learn English collocations and five tools used to learn Spanish collocations before we developed our computer-assisted collocation learning tool for Spanish. A summary of the features and disadvantages of each tool is provided below.

With regard to English assisted learning tools, a POS (part of speech) search is not available in the Hong Kong Polytechnic Web Concordancer (Greaves, 1999), while in TANGO (Jian et al., 2004), the POS of a keyword can be defined, and example and frequency are also provided, but types of POS are limited. In WebCollocate (Chen, 2011), a POS search is available, and search results are sorted by frequency, with a user-friendly interface and the provision of related sentences, but it is not currently available for public use. In addition, there are also collocation assisted learning tools for bilingual uses such as TOTALrecall (Wu et al., 2003), which can be searched in both Chinese and English. In Writing Assistant (Chang et al., 2008), user mistakes can be revised with the correct collocation based on Chinese-English translations. Furthermore, English collocation assisted learning tools provide customized search functions such as the Corpora and NLP for Digital Learning of English, CANDLE (Liou et al., 2006), which consists of three sub-systems tailored to different user levels; the Writing-Collocation Checker, which can automatically detect "verb + noun" collocations and provide correct collocations for users, and Linggle (Boisson et al., 2013), which has selective preference and synonym group functions and provides different types of arguments based on the predicate.

With respect to Spanish collocation assisted learning tools, CrossLexica Española (Bolshakov & Miranda-Jiménez, 2004) is a Spanish collocation assisted learning tool with a POS search function, and it is probability-based, with a grammatical function and semantic classification available, but it is not available for public use. The Corpus del Español, CdE (Davies, 2012) is more advanced, with lemmas functions, and is user friendly, but it provides too many examples and may be difficult for beginners to use. Diccionario de Colocaciones del Español, DiCE (Alonso Ramos et al., 2010) is free for users and provides general and advanced functions for searching for collocations through lexical lemma entities on specific themes, such as emotion nouns (for example, alegría "joy" and estima "esteem") with semantic "feeling" and "mental actions" features such as "sentir una gran alegría" or "alta estima". Syntactical structures, meaning identi-

fication, explanations, and examples associated with a list of lexical units are included to illustrate the searched collocations. The DiCE is a powerful online dictionary in terms of providing lexical information, but only a Spanish interface is available, so its high-level collocation might be difficult to understand for learners with limited proficiency in Spanish.

In addition, Sketch Engine (Kilgarriff et al., 2014) is a Spanish collocation assisted learning tool with multilingual search functions. It can select different statistical methods according to language features, but the statistical results are relatively complicated. Finally, EuroWordNet (González-Agirre & Rigau, 2013) includes a variety of European languages such as English, Spanish, and Italian, but the search results are research-oriented and might be too advanced and complicated for foreign language learners to understand and apply, especially in the case of those who are at the beginning and intermediate levels.

2.2 Evaluation of collocation assisted learning tools

In a review of the studies related to an evaluation of the developed assisted tools, it was found that Chen (2011) investigated the relative effectiveness of several computer assisted English collocation tools focusing on two groups of users, learners and teachers of English. Students from two similar classes used different tools to translate sentences from Chinese into English, whereas English teachers assessed four English collocation learning tools. The results showed that students who used WebCollocate (the developed assisted learning tool in English by Chen (2011)) performed better than those who used the other tool, Hong Kong Polytechnix Web Concordancer. Language teachers reported that using WebCollocate was less time consuming and that it was easier to search for collocations and to find many collocation examples because of the large database in the corpus.

With respect to Spanish collocation tools, Vincze et al. (2011) extended a series of collocation-related analyses based on DiCE (Diccionario de Colocaciones del Español) to studies of computer-assisted language learning; the authors utilized CEDEL2, an L1 English-L2 Spanish learner corpus (Lozano, 2009), to develop a computer-assisted learning tool for Spanish collocations. Alonso Ramos et al. (2010) annotated both the correct and incorrect collocations in the learner corpus to find collocations undetected by auto-correction tools with an analysis of error features, so as to improve the error-detection function of the collocation learning tool. Their analysis of collocation errors included recognizing collocations, correction judgment and interpretation of errors. Ferraro et al. (2014) pointed out that there are only a few tools that provide users with high accuracy and proper corrections, and most tools only offer a list of collocation options for users to choose from. For the detection of incorrect collocations, Ferraro et al. (2014) employed frequency-based techniques and attempted to provide users with proper corrections rather than simply listing all the possible corrections. They argued that although ordered lists might be helpful for advanced learners, the tool would not be as beneficial for learners at the elementary and intermediate levels, especially when the suggested lists include words with subtle semantic differences that are difficult to distinguish one from the other.

2.3 Acquisition of Spanish collocation

Among the available research on the acquisition of Spanish collocation, Laufer & Waldman (2011) found that learners at different proficiency levels used fewer collocations than native speakers. Previous studies also showed that collocation causes various degrees of difficulty for learners from beginning to advanced levels in the lexical learning process. With regard to different types of collocations, previous research (Laufer & Waldman, 2011; Nesselhauf, 2003; Alfahadi et al., 2014) has concentrated more on the adjective-noun (AdjN) and the verb-noun (VN) constructions, which are considered more problematic for learners. Going one step further, Lu & Cheng (2016) compared and contrasted four different essential types of Spanish combinations, VN, AdjN, NAdj, and VP in learner and parallel corpora. The results showed a sequence of development from NAdj, VN, to AdjN combinations. The results also suggested that most learner errors were related to the learners' L1 (Chinese) and L2 (English). Furthermore, lexical errors might be associated with the form-meaning transfer from the previous languages of learners.

As in the aforementioned learning tools for Spanish collocations intended to extend related studies, in this research, built upon previously constructed corpora, a computer-assisted learning system was developed with two major functions, error detection and revision suggestions for Spanish collocation, and an experiment was conducted in order to evaluate its effectiveness in terms of learning.

3 Research method

The methodology involved in this study included two major parts. The first one was the development of a corpus-based learning tool for Spanish collocation, and the second part was an evaluation of the developed learning tool. Based on the previous development experience using SEDRS (System of Error Detection and Revision Suggestion), the construction of the Spanish Collocation Error Detection and Revision Suggestion tool (SpColEDRS) involved the employment of data sources from two corpora (the Corpus Escrito de Aprendices Españoles / Learners' Written Corpus of Spanish, CEATE and the Corpus Paralelo de Español, Inglés y Chino / Parallel Corpus of Spanish, English and Chinese, CPEIC) and a data analysis and collocation extraction tool (Herramienta de Colocación Española / Spanish Collocation Tool, HCE). After developing the computer-assisted learning tool, SpColEDRS, with two major functions (error detection and revision suggestions), an experiment was conducted and a questionnaire was used to evaluate its effectiveness for checking Spanish collocations from the perspective of learners.

3.1 The development of a computer-assisted learning tool: SpColEDRS

The first part of this section addressed the development of the assisted learning tool, the Spanish Collocation Error Detection and Revision Suggestion (SpColEDRS). Texts were analyzed and processed using the POS tagging system, and then collocations were calculated and extracted as outputs through the Spanish collocation tool (HCE) search functions. To extract collocations from the data source, a statistical method was employed. It was defined so as to test whether the probability of two co-occurring elements in a combination was under the confidence level. Based on a highly-cited study by Manning et al. (1999), χ^2 (or Chi-squared) was determined as the statistical method for the extraction of collocations used to develop the assisted learning tool, SpColEDRS. The training data (9,807 words) for the developed tool, comprised the fairy tales[1] from the Spanish subcorpus of the CPEIC trilingual parallel corpus and revised texts from the CEATE learners' corpus. The database of Spanish collocations was generated with machine learning and processed through data processing, collocation extraction, and manual modification. This database served as a reference to carry out collocation checking by detecting learner errors and providing possible suggestions for learners to use to correct their errors. TreeTagger was used for POS-tagging data, and PHP, AJAX, and MySQL were used as the development tools for error detection and revision suggestions. The SpColEDRS tool was designed with two main functions: error detection and revision suggestions for Spanish collocation for learning purposes.

[1] Data sources included International Children's Library http://en.childrenslibrary.org/ and http://itunes.apple.com/hk/app/id440153337?mt=8.

3.2 The evaluation of the computer-assisted learning tool for Spanish collocations

To evaluate the practical effectiveness of the developed tool from the user perspective, we conducted an experiment consisting of a pretest, a video tutorial, a post-test, followed by a user questionnaire. The collected information was analyzed to examine whether the SpColEDRS tool was able to assist learners with improving their learning by comparing learning outcomes from two groups of Spanish learners, experimental and contrastive groups.

3.2.1 Participants

Thirty three (33) Spanish learners from National Cheng Kung University participated in the evaluation. Their mother language was Mandarin-Chinese; their first foreign language (L2) was English, and their second foreign language (L3) was Spanish, in which they had 180–360 instructed hours. The participants did not have much contact with the L3 Spanish outside of the classroom since Mandarin Chinese is the predominant language in Taiwan. Prior to the pretest, all participants took the Wisconsin Placement Test to assess their Spanish proficiency in general. According to their scores on the Wisconsin Placement Test, they were grouped into two proficiency levels of Spanish: 11 at beginning-high level (457-517 points) and 22 at the intermediate-low levels (535-653 points). Then, they were randomly assigned to two groups, 17 to the experimental and 16 to the contrastive groups.

3.2.2 Procedure

Both the on-line pre and post-tests contained 40 sentences with one element of the combination left blank to be filled in by the participants according to the correspondent translation in Chinese (Appendix 5). The tested combinations were four different types, including Verb-

Noun, Adjective-Noun, Noun-Adjective, and Verb-Preposition.

One week after the pretest was conducted, the participants in the experimental group were directed to view a video tutorial (two-minutes) to learn how to use the SpColEDRS computer-assisted learning tool. The participants in the control group did not receive any treatment. The video tutorial provided participants with basic instructions for using the assisted learning tool.

Then, the participants in both groups completed the post-test engaging in the same task as that used in the pretest. In the post-test, the participants from the experimental group were required to fill in a blank to complete the combined elements of the collocation before using the assisted tool, and then on another line, they were asked to indicate whether they modified the answer after using the provided tool and to explain what they had changed if this was the case (Appendix 5).

After the post-test, the participants from the experimental group were required to complete the questionnaire (Appendix 5). The questionnaire included two subsections; one was used to collect the users' levels of satisfaction with the interface on a Likert-scale, and the other involved open-ended questions regarding the usefulness of the system as well as suggestions for further modifications.

4 Results and discussion

4.1 Development of SpColEDRS

The developed computer-assisted learning tool for Spanish collocation provides a checking functionality with error detection and revision suggestions for Spanish collocation, as shown in Figure 1. If the key-in collocation exists in our database, the system responds with a confirmation, as shown in Figure 2. However, when a possible error is entered, the system responds immediately, and users can then select an appropriate revision from the provided suggestion list, as shown in Figure 3.

Figure 1: Spanish Collocation Error Detection and Revision Suggestion tool: User's interface.

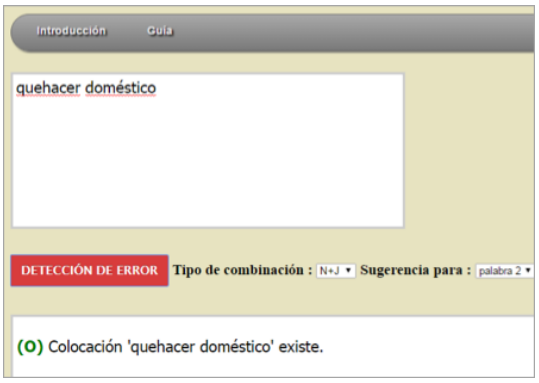

Figure 2: Spanish Collocation Error Detection and Revision Suggestion tool: Confirmation of correct use.

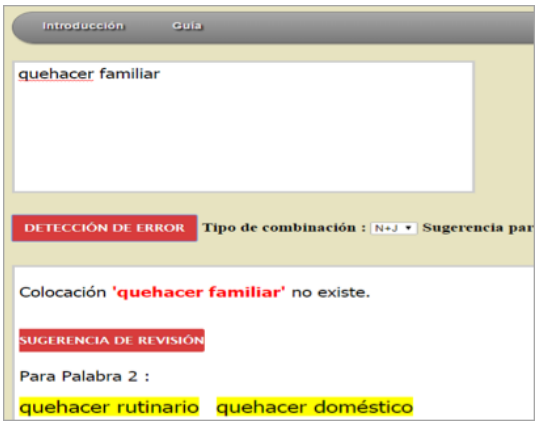

Figure 3: Spanish Collocation Error Detection and Revision Suggestion tool: Provision of correction suggestion.

4.2 Evaluation of developed tool

4.2.1 Data Analysis Methods

According to the research questions, a one-way ANCOVA was selected for the purpose of determining (1) if there were any significant differences between the experimental and control groups after the pedagogical intervention using the developed corpus-based learning tool, and (2) if there were any significant differences between learners at the beginning and intermediate levels in the post-test after using the assisted learning tool.

Prior to the analysis of research question 1, the pretest, post-test, and group variables were examined using SPSS programs to check for the accuracy of data entry, missing values, the linearity between the covariate (pretest) and dependent variables (post-test), and the assumptions of the homogeneity of the regression slopes, normality, homoscedasticity, homogeneity of variance, and outliers.

There were no missing values in the data set. Pairwise linearity was checked using within-

group scatterplots and was found to be satisfactory. There were no cases detected as outliers based on an examination of the z scores on the post-test. There was homogeneity of the regression slopes because the interaction term was not statistically significant, $F(1, 29) = 0.000$, $p = 0.982$. Because the variable post-test was severely skewed, a "reflect and logarithmic" transformation was applied, which means that the new post-test variable (PostTotal-log10-ref) was equal to the LG10 ("the highest score on the post-test plus 1" — post-test scores"). With the transformed variable in the variable set, standardized residuals for the post-test and for the overall model were normally distributed, as assessed with the Shapiro-Wilk's test ($p > 0.05$). Also, there was homoscedasticity, as assessed by visual inspection of the standardized residuals plotted against the predicted values. The assumption of homogeneity of variances was met, as assessed by Levene's test of homogeneity of variance ($p = 0.246$). There was no outlier in the data, which was assessed by determining that there were no cases with standardized residuals greater than ±3 standard deviations.

4.2.2 Effectiveness

A one-way ANCOVA was run to determine the effect of the pedagogical intervention treatment using the corpus-based learning tool developed for this study on the post-test after controlling for the pretest. As shown in Table 1, after adjustment for the pretest, there was a statistically significant between-group difference in the post-test for the experimental group and the control group, $F(1, 30) = 100.768$, $p < 0.001$, partial $n^2 = 0.771$.

Source	*df*	*MS*	*F*	*p*	n^2
PreTotal	1	0.219	7.189	0.012	0.193
Group	1	3.069	100.768	0.000	0.771
Error	30	0.030			

Table 1: Analysis of covariance for the post-test with the pretest as a covariate.

The post hoc analysis was performed with a Bonferroni adjustment. The post-test scores were statistically significantly better in the experimental group than in the control group, as shown in Tables 2 and 3, because the post-test scores were transformed by a "reflect and logarithmic" transformation as explained above. Therefore, the developed assisted learning tool had a positive effect on the students' learning of Spanish collocations.

		Unadjusted		Adjusted	
	N	M	SD	M	SE
Control	16	14.062	4.3277	13.8424	0.8132
Experiment	17	24.118	2.5952	24.3248	0.7887

Table 2: Adjusted and unadjusted experimental means and variability for the post-test with the pretest as a covariate before post-test transformation.[2]

		Mean			95% C.I.$^{(b)}$	
Group (I)	Group (J)	Difference (I-J)	SE	Sig.$^{(b)}$	Lower Bound	Upper Bound
C	E	0.616$^{(a)}$	0.061	0.000	0.491	0.742
E	C	-0.616$^{(a)}$	0.061	0.000	-0.742	-0.491

Table 3: Pairwise comparison.[3]

In addition, a one-way ANCOVA was also selected to answer research question 2: Are there any significant differences between the different levels of learner proficiency in the post-test after using the learning tool? The same statistical analysis procedures used for research question 1 were conducted. The variable post-test was also transformed using a "reflect and logarithmic" transformation as explained above because the post-test variable has a serious skewness. With the transformation, all the assumptions for the one-way ANCOVA were satisfied.

The results of the one-way ANCOVA test shown in Table 4 show that there was no significant difference between the beginning level and intermediate level in the post-test after using the learning tool by controlling the effect of pretest, $F(1, 30) = 0.088$, $p = 0.769$.

Source	*df*	*MS*	*F*	*p*	n^2
PreTotal	1	0.016	0.120	0.732	0.004
Level	1	0.012	0.088	0.769	0.003
Error	30	0.132			

Table 4: Analysis of covariance for the post-test with the pretest as a covariate.

However, the results of the independent *t*-test shows that there was a significant difference between the beginning level and intermediate level in the pretest before using the learning tool, $p < 0.001$. After using the learning tool, the

[2] N is the number of participants, M the Mean, SD the Standard Deviation and SE the Standard Error.

[3] Dependent Variable: PostTotal-log10-ref; Based on estimated marginal means; $^{(a)}$ The mean difference is significant at the .05 level.; $^{(b)}$ Adjustment for multiple comparisons: Bonferroni.

beginning group increased their test scores from 7.909 to 18 (see Table 5). Also, the intermediate group increased their test scores from 13.0909 to 19.8636 (see Table 5). Therefore, the learning tool had a positive effect on both the beginning group and the intermediate group, but had a greater positive effect on the beginning group.

		Pretest		Post-test	
		M	SD	M	SD
Beginning	11	7.909	3.113	18	7.4027
Intermediate	22	13.0909	4.0344	19.8636	5.5574

Table 5: Means and variability for the pretest, and the post-test with the pretest as a covariate.[4]

4.2.3 Questionnaire

The results of the satisfaction survey for the interface interaction between the users and the developed tool showed that most participants were satisfied (over 3.8 on a scale of 5) with the SpColEDRS in terms of identifying collocation errors and the suggestion lists provided to them for correction, as shown in Table 6. According to the user responses, the assisted learning tool was easy and simple to use, and the reaction times for error detection and correction suggestions for Spanish collocations were appropriate. This developed tool was recommended for self-learning although users at different proficiency levels might benefit from it to a greater or lesser degree. In summary, the Spanish collocation error detection and correction suggestion functions for the lexical features included in the database were found to be useful.

Interface interaction	Q1	Q2	Q3	Q4	Q5
Beginning	4.5	3.8	4.8	5	4.5
Intermediate	4.5	4.3	4.1	4.5	4.5

Table 6: Results indicating satisfaction with the Spanish Collocation Error Detection and Revision Suggestion tool[5].

According to the participants' responses to the open-ended questions in the survey, the advantages of this collocation learning tool included immediate feedback and ease and simplicity of the search process. However, the tool had several disadvantages. For example, users had to know at least one word of the two combined elements in order to make it possible to use the tool. It was difficult to choose the appropriate one from more than one possible correction suggestion. The users suggested future modifications such as to provide English or Chinese translations of the searched collocations to facilitate understanding of the meaning of the collocations. The participants also suggested providing examples of collocation usage to help distinguish subtle differences among the collocations offered in the feedback.

4.3 Limitations and future work

The user evaluation of the SpColEDRS was, in general, positive and suggested that the users were satisfied. However, the training data for our developed tool from the two corpora (learners' corpus CEATE and trilingual parallel corpus CPEIC) was relatively small. Therefore, the identification and detection of errors were limited to collocations within a fixed and limited range. Also, the context and the current experiment were conducted within searchable combinations. A larger amount of training data from a greater variety of text types should be included for training in the future in order to obtain better results in terms of error detection and correction suggestions, which would strengthen the applicability of this assisted corpus-based tool for teaching and learning Spanish collocations.

As has been suggested by users, translations of L1 Chinese or L2 English should be provided to assist learners with their understanding of Spanish collocations, especially in the case of beginning learners. In addition, examples of collocation uses in sentences in meaningful contexts should be listed as an option to illustrate the differences among the suggested collocations.

5 Conclusions

In this study, a corpus-based assisted tool for collocation learning in Spanish was developed and evaluated. Based on the training data compiled in two created corpora (CEATE and CPEIC) and a Spanish collocation extraction tool (HCE), this computer-assisted learning tool is easy to operate and has two major functions: error detection and revision suggestions. SpColEDRS can detect inappropriate uses of Spanish collocations and provides suggestion lists for learners to choose from for the purpose of correcting their collocation errors.

[4] N is the number of participants, M the Mean and SD the Standard Deviation.

[5] Q1: Identifying lexical errors; Q2: Provision of suggestion list for correction; Q3: Easy and simple to use; Q4: Appropriate reaction time; Q5: Recommended for self-learning).

To ensure the effectiveness of and user satisfaction with the SpColEDRS, the developed tool was evaluated using two tests and a questionnaire. The research results showed that the SpColEDRS could assist learners effectively based on the progress of the experimental group from the pretest to the post-test, especially in the case of the beginning learners. Furthermore, the results of the satisfaction survey assessing the students' opinions of the interface and usefulness of the tool indicated that most of the participants positively confirmed that the tool was effective for assisting them with their practice with Spanish collocations. Finally, to optimize the use of the existing corpora (CEATE and CPEIC) and tool (HCE), this study extended our previous outcomes of the created corpora and tool for the advancement of studying effective learning of Spanish collocation in Taiwan and further shed light on pedagogical applications of the created corpora and on the learning of Spanish collocation with a corpus-based approach in a multilingual acquisition setting.

Acknowledgments

We wish to extend our sincere gratitude to the Ministry of Science and Technology of Taiwan for their generous support with project grant number 103-2410-H-006-059-MY2, and our appreciation for the technical support provided by the Computer Science and Information Engineering team at National Cheng Kung University in Taiwan and the research assistants involved in this project.

References

Alfahadi, Abdulrahman M., Said Ahmed Zohairy, Mowaffaq Mohammed Momani & Mansour H. Wahby. 2014. Promoting awareness of teaching collocations techniques to beginners (adjective-noun collocations). *European Scientific Journal* 10(10). 389–396.

Alonso Ramos, Margarita, Alfonso Nishikawa & Orsolya Vincze. 2010. DiCE in the web: An online spanish collocation dictionary. In *ELexicography in the 21st Century: New Challenges, New Applications (ELex)*, 369–374.

Boisson, Joanne, Ting-Hui Kao, Jian-Cheng Wu, Tzu-Hsi Yen & Jason S. Chang. 2013. Linggle: a web-scale linguistic search engine for words in context. In *51st Annual Meeting of the Association for Computational Linguistics*, 139–144.

Bolshakov, Igor A. & Sabino Miranda-Jiménez. 2004. A small system storing Spanish collocations. In *International Conference on Intelligent Text Processing and Computational Linguistics*, 248–252.

Chang, Yu-Chia, Jason S. Chang, Hao-Jan Chen & Hsien-Chin Liou. 2008. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning* 21(3). 283–299. 10.1080/09588220802090337.

Chen, Hao-Jan Howard. 2011. Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning* 24(1). 59–76. 10.1080/09588221.2010.526945.

Davies, Mark. 2012. Corpus del español (100 million words, 1200s-1900s). [online] http://www.corpusdelespanol.org.

Ferraro, Gabriela, Rogelio Nazar, Margarita Alonso Ramos & Leo Wanner. 2014. Towards advanced collocation error correction in Spanish learner corpora. *Language Resources and Evaluation* 48(1). 45–64. 10.1007/s10579-013-9242-3.

Firth, John. 1957. Modes of meaning. *Papers in Linguistics* 5. 190–215.

González-Agirre, Aitor & German Rigau. 2013. Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. *Linguamática* 5(1). 13–28.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology: Theory, analysis, and applications*, 145–160. Oxford University Press.

Greaves, Christopher. 1999. Virtual language centre study guide. [online] http://vlc.polyu.edu.hk/.

Jian, Jia-Yan, Yu-Chia Chang & Jason S. Chang. 2004. TANGO: bilingual collocational concordancer. In *Association for Computational Linguistics (ACL)*, 10.3115/1219044.1219063.

Källkvist, Marie. 1995. Lexical errors among verbs: A pilot-study of the written language of advanced Swedish learners of English. *Working Papers in English and Applied Linguistics* 103–115.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1). 7–36. 10.1007/s40607-014-0009-9.

Laufer, Batia & Tina Waldman. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61(2). 647–672. 10.1111/j.1467-9922.2010.00621.x.

Liou, Hsien-Chin, Jason S Chang, Hao-Jan Chen, Chih-Cheng Lin, Meei-Ling Liaw, Zhaoming Gao, Jyh-Shing Roger Jang, Yuli Yeh, Thomas C. Chuang & Geeng-Neng You. 2006. Corpora processing and computational scaffolding for a web-based English learning environment: The CANDLE project. *CALICO journal* 24(1). 77–95.

Lorenz, Gunter R. 1999. *Adjective intensification: learners versus native speakers: a corpus study of argumentative writing*, vol. 27. Rodopi.

Lozano, Cristóbal. 2009. CEDEL2: Corpus escrito del español L2. In Bretones Callejas (ed.), *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente.*, 197–212. Universidad de Almería.

Lu, Hui-Chuan & An Chung Cheng. 2016. Acquisition of L3 Spanish combinations: Development in bilingual and multilingual contexts. In *8th International Conference of Language Acquisition*, n.pp.

Lu, Hui-Chuan, Yu-Hsin Chu & Cheng-Yu Chang. 2013. A corpus-based system of error detection and revision suggestion for spanish learners in taiwan: A case study. *JALT CALL Journal* 9(2). 115–130.

Manning, Christopher D, Christopher D Manning & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

McKeown, Kathleen R. & Dragomir R. Radev. 2000. Collocations. In *Handbook of Natural Language Processing*, 1–23. CRC Press.

Nattinger, James R. & Jeanette S. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford University Press.

Nesselhauf, Nadja. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics* 24(2). 223–242. 10.1093/applin/24.2.223.

Nesselhauf, Nadja. 2005. *Collocations in a learner corpus*, vol. 14. John Benjamins Publishing.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Vincze, Orsolya, Margarita Alonso Ramos, Estela Mosqueira Suárez & Sabela Prieto González. 2011. Exploiting a learner corpus for the development of a CALL environment for learning Spanish collocations. In *Electronic lexicography in the 21st century: New applications for new users (eLex)*, 280–285.

Weisser, Martin. 2016. Corpus-based linguistics links. [online] http://martinweisser.org/corpora_site/CBLLinks.html.

Wray, Alison. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied linguistics* 21(4). 463–489. 10.1093/applin/21.4.463.

Wu, Jian-Cheng, Kevin C. Yeh, Thomas C. Chuang, Wen-Chi Shei & Jason S. Chang. 2003. TOTALrecall: A bilingual concordance for computer assisted translation and language learning. In *41st Annual Meeting on Association for Computational Linguistics*, 201–204. 10.3115/1075178.1075216.

Appendices

Appendix 1-Control and experimental group pretest and control group post-test.

1 我現在很累，想睡覺。
Estoy muy cansado.___sueño.

2 那時我們看著101的煙火，在美國餐廳吃晚餐。
Vimos los fuegos__ del 101 y cenamos en un restaurante americano.

3 她留有一頭長且直的金髮。
Tiene el pelo___, largo y liso.

4 在我空閒的時間裡，我喜歡聽音樂。
En mi ___libre, me gusta escuchar la música.

5 現在我和我的父母關係很好，因為我試著和他們溝通。
Ahora tengo unas buenas relaciones con mis padres porque yo trato ___ comunicarme con ellos.

6 他們生氣不是沒有道理的，您是應該專心讀書的。
Ellos no están enojados sin razón, usted tiene que ___atención al estudiar.

7 第二天，我們參觀一間很大的博物館。
Al día___, visitamos un museo muy grande.

8 他突然因為撞到了一塊石頭而跌倒了。
De pronto tropezó___ una piedra y se cayó.

9 你們必須知道父母將不會對子女造成傷害。
Necesitáis saber que los padres no van a ___daño a sus hijos.

10 他不僅編新故事，還找藉口。
Inventaba nuevas historias y ___excusas.

11 哥斯大黎加的傳統經濟依賴農業。
La economía tradicional de Costa Rica depende___ la agricultura.

12 星期一到五下課後，我習慣悠遊於網際網路之中。
De lunes a viernes, suelo navegar___ internet después de terminar la clase.

13 通常夏天都是很好的天氣，可是有時會下很大的雨，而且有風。
Generalmente, hace buen tiempo, pero a veces llueve mucho y ___viento en verano.

14 女人為了漂亮，有很多方法可以減重。
Para que las mujeres sean bonitas hay muchos métodos para ___peso.

15 他無法支撐起他的家庭，而且還有一大堆的問題。
No puede sostener___ su familia y tiene una cantidad de problemas.

16 我有可能環遊全世界、認識各式各樣的人。
Tengo la posibilidad para viajar en el mundo___ y conozco toda clase de personas.

17 他是很用功的學生，上課總會問問題。
Es un estudiante muy trabajador. Siempre ___preguntas en clase.

18 因為想家，我剛才打電話給我母親。
Debido a la nostalgia, acabo ___ llamar a mi madre por teléfono.

19 為了買到夢想中的房子，他努力省錢。
Para comprar la casa de sueño, hace un esfuerzo por ___dinero.

20 我的弟弟今年將要上小學。
Mi hermano menor va a ir a la escuela____ este año.

21 如果他想看電視，需要先完成作業。
Si quiere ver la televisión, ___falta terminar la tarea primero.

22 他沒有兄弟姊妹，是獨生子。
No tiene hermanos, es hijo ___.

23 這個故事發生在一個小村莊。
La historia ___lugar en un pequeño pueblo.

24 我們家有四間臥室和二間廁所。
Nuestra casa cuenta _ cuatro dormitorios y dos baños.

25 然後，那些今年畢業生排隊離開？
Luego los estudiantes que gaduarán este año ____cola y salieron del auditórium.

26 如果您的小孩犯錯的話，不要馬上處罰他們。
Si sus hijos ____ errores, no les castigue inmediatamente.

27 夜市中有很多美食。
Hay muchas comidas deliciosas en el mercado nocturno_.

28 你必須小心不要踩到草皮，不然會被罰款。
Tienes que _ cuidado de no pisar el césped, te ponen una multa.

29 我一直讀一直讀，終於覺得那些文字開始合乎邏輯了。
Continué leyendo y leyendo, y finalmente las palabras comenzaron a ___sentido.

30 他不喜歡幫助別人，也從不向人求助。
No le gustaba ayudar a otros y nunca ___ayuda a otros.

31 跟我談談你下一年度的計畫吧！
Dime algo sobre tus planes para el ____año.

32 這個節目提供了一些安全騎機車的資訊和建議。
Este programa proporciona los informaciones y consejos para andar ____ motocicleta con seguridad.

33 為了獲取好成績，我很用功讀書。
Para ____nota buena, estudio mucho en la biblioteca.

34 我很欣賞他作詩的才華。
Le admira mucho por su capacidad para ___versos

35 他很喜歡開玩笑，因此有很多朋友。
Le gusta ___broma, por eso, tiene mucho amigos.

36 我想要跟那位電影明星一起合照。
Quiero que ___foto con esta estrella de cine.

37 我想要享受一個美好的退休生活。
Quisiera disfrutar ___ una vida maravillosa después de la jubilación.

38 因為我會餓，所以我想快點結束。
Quiero acabar rápido porque ___hambre.

39 因為他奇怪的笑容讓我們害怕，所以我們開始像瘋子一樣跑了起來。
Empezamos a correr como locas porque ___miedo su extraña sonrisa.

40 我認為我們必須注意/理會醫生的建議。
Creo que deberíamos ___caso a los consejos del doctor.

Appendix 2-Pretest of control and experiment groups, and posttest of control group.

1 我現在很累，想睡覺。
- a. Estoy muy cansado. _____sueño.
- b. 使用EDRS後未修正。修正為：_____sueño

2 那時我們看著101的煙火，在美國餐廳吃晚餐。
- a. Vimos los fuegos _____ del 101 y cenamos en un restaurante americano.
- b. 使用EDRS後未修正。修正為：fuegos _____

3 她留有一頭長且直的金髮。
- a. Tiene el pelo _____, largo y liso.
- b. 使用EDRS後未修正。修正為：pelo _____

4 在我空閒的時間裡，我喜歡聽音樂。
- a. En mi _____ libre, me gusta escuchar la música.
- b. 使用EDRS後未修正。修正為：_____ libre

5 現在我和我的父母關係很好，因為我試著和他們溝通。
- a. Ahora tengo unas buenas relaciones con mis padres porque yo trato _____ comunicarme con ellos.
- b. 使用EDRS後未修正。修正為：trato _____

6 ∼ 40 Chinese...
- a. Spanish sentences...
- c. 使用EDRS後未修正。修正為：...

Appendix 3-Questionnaire

SpColEDRS: Questionnaire
「工具」使用意見調查
Please indicate if you agree or disagree with the following statements by using the scale from 5 (strongly agree) to 1 (strongly disagree.)
□５：非常同意、□４：同意、□３：普通、□２：不同意、□１：非常不同意
□5: Strongly agree, □4: Agree, □3: Neutral, □2: Disagree, □1: Strongly disagree

1. System interface
一、系統之操作介面互動
• SpColEDRS can detect errors in lexical usage?
能偵測出詞彙使用錯誤
□５□４□３□２□１
• SpColEDRS can provide helpful suggestions for revision?
所建議的清單對修正有幫助
□５□４□３□２□１
• SpColEDRS is easy to use?
容易操作
□５□４□３□２□１
• SpColEDRS can detect errors within an acceptable response time?
系統偵測錯誤所需的時間適當
□５□４□３□２□１
• SpColEDRS can provide suggestions for revision within an acceptable response time?
系統給予建議修正所需的時間適當
□５□４□３□２□１
• In general, SpColEDRS facilitates the self-learning of Spanish in writing?
整體而言，對西語表達的自學有幫助
□５□４□３□２□１

2. Advantages and disadvantages of SpColEDRS
二、「錯誤偵測、修正建議系統」的優點及缺點
Based on your user experience, please evaluate the SpColEDRS regarding the advantages and disadvantages of the following features.
(1) Error detection
(一)「錯誤偵測」部分：
Advantages 優點：

Disadvantages 缺點：
---
(2) Revision suggestion
(二)「修正建議」部分：
Advantages 優點：

Disadvantages 缺點：
---

3. Feedback and suggestions for SpColEDRS
三、對「錯誤偵測、修正建議系統」之建議及回饋
Based on your user experience with the SpCoIEDRS, please provide your feedback and suggestions for improvement in the following areas:
(1) Error detection
(一)「錯誤偵測」部分：
---
(2) Revision suggestion
(二)「修正建議」部分：

Thank you for your cooperation!
謝謝配合！

Projetos, Apresentam-se!

Propuesta recibida el octubre 2018 y aceptada para publicación el junio 2019.

SAUTEE: un recurso en línea para análisis estilométricos

SAUTEE: an online resource for stylometric analysis

Fernanda López-Escobedo
Universidad Nacional Autónoma de México
flopeze@unam.mx

Gerardo Sierra
Universidad Nacional Autónoma de México
gsierram@iingen.unam.mx

Julián Solórzano
Universidad Nacional Autónoma de México
jsolorzanos@iingen.unam.mx

Resumen

La estilometría es la cuantificación del estilo por medio de la búsqueda de rasgos textuales que sean medibles y representativos del estilo de un autor. No existen muchas aplicaciones dirigidas al público en general que permitan realizar estudios de esta naturaleza, y las que existen son relativamente limitadas o no necesariamente amigables al usuario. En este artículo presentamos una aplicación web para análisis estilométrico. La aplicación está respaldada por un gestor de corpus, es de fácil manejo y presenta los resultados de manera intuitiva, sin dejar de lado la visión de ofrecer un catálogo exhaustivo de marcadores estilométricos y métodos de análisis.

Palabras clave

estilometría, atribución de autoría, lingüística forense

Abstract

Stylometry is a method that quantifies writing styles by isolating and counting distinctive and measurable textual features of an individual's style. Currently, there are few software applications, aimed at a wide user base, capable of performing such an analysis. Of those readily available, most suffer from limited computing power and/or are not user-friendly. In contrast, our web-based stylometric application, backed by a robust corpus manager, is easy to use, offers a thorough catalogue of stylometric markers and analytic methods to choose from, and produces an intuitive readout of the results.

Keywords

stylometry, authorship attribution, forensic linguistics

1 Introducción

Tradicionalmente se han aplicado los estudios del estilo literario a problemas cronológicos y obras de autoría disputada, como por ejemplo el caso de las obras de Shakespeare. Desde finales del siglo XIX se han intentado establecer métodos numéricos y estadísticos que permitan medir el estilo de un autor y hasta hace poco esta tarea era vista principalmente como una ayuda complementaria en estudios de humanidades. Sin embargo, desde la segunda mitad del siglo XX estas ideas empezaron a ser de interés para el ámbito legal, por ejemplo véase Svartvik (1968), en donde se usaron técnicas estadísticas para demostrar que las supuestas confesiones de Timothy John Evans —hombre acusado de asesinar a su esposa e hija y condenado a muerte en 1950— fueron alteradas por la policía. Esta evidencia fue importante para su perdón póstumo (Nieto et al., 2008). En los 90's el lingüista Malcom Coulthard dio forma a la subdisciplina que hoy en día se conoce como lingüística forense. El testimonio presentado por Coulthard en el caso de otro hombre acusado de asesinato, Derek Bentley, demostró, similar a lo ocurrido en el caso Evans, que la supuesta confesión grabada había sido fabricada (Coulthard, 1994).

La estilometría es una línea de investigación dentro del ámbito de la lingüística forense que tiene como objetivo cuantificar el estilo o, en otras palabras, analizarlo estadísticamente. Para ello se busca identificar ciertos rasgos que sean comunes en el lenguaje pero que sean característicos de cada autor. Es decir, se basa en el supuesto de que hay un factor inconsciente, pero distintivo y medible, en el estilo de escritura de cada persona (Holmes, 1998).

Se considera que uno de los primeros trabajos en este ámbito es el realizado por el físico Thomas Mendenhall en 1887, quien analizó en las obras de Shakespeare la distribución de la frecuencia de las palabras de acuerdo con su longitud (Mendenhall, 1887). Aunque no demostró nada contundente, surgió un interés por el te-

DOI: 10.21814/lm.11.1.270
This work is Licensed under a Creative Commons Attribution 4.0 License

ma y en las décadas subsecuentes se propusieron otros marcadores estilométricos, con éxito moderado. Más tarde, en 1964, el trabajo de Mollester y Wallace acerca de la autoría de "The Federalist Papers" (Mosteller & Wallace, 1964) se posicionó como un parteaguas en el área debido a sus convincentes resultados y su entonces novedosa técnica Bayesiana. En los años posteriores se adoptaron técnicas de estadística multivariada y varios tipos de análisis usando algoritmos de aprendizaje de máquina (*machine learning*), como por ejemplo máquinas de soporte de vectores (Diederich et al., 2003) y redes neuronales artificiales (Tweedie et al., 1996).

La estilometría y la atribución de autoría siguen siendo objeto de estudio y polémica, ya que no se ha podido definir un conjunto de marcadores estilométricos universales que puedan consistentemente identificar a cualquier autor en cualquier situación; ni tampoco se ha aceptado una metodología universal para llevar a cabo una tarea de atribución de autoría en el ámbito forense. Hasta el día de hoy se lucha por encontrar un protocolo estándar, por ejemplo véase la reciente propuesta de Juola (2015).

Con el afán de incrementar la investigación en el área y hacerla más conocida, es deseable contar con herramientas utilizables por usuarios finales que les permitan conocer y evaluar las diversas técnicas existentes. Esto es, por ejemplo, acercar al área a lingüistas que no necesariamente están familiarizados con la estadística o la computación, a profesionistas del ámbito legal que deseen evaluar la confiabilidad de los resultados, y otras personas interesadas que no sean expertos en análisis cuantitativo. De hecho, actualmente no existen muchas opciones de software que cumplan con estas características, por lo menos no para el uso del público en general.

El presente artículo tiene como objetivo presentar un nuevo sistema que cumple con las características de ser amigable al usuario sin dejar de lado el poder de sus análisis. Además, tiene la particularidad de ser una aplicación web, con miras a explotar ventajas como son el hecho de no requerir instalación, de poder usarse desde cualquier computadora en cualquier momento y de estar respaldado por un gestor de corpus colaborativo.

En la siguiente sección se presentan las herramientas existentes más conocidas, discutiendo brevemente sus ventajas y desventajas. En la Sección 3 se describe el marco del proyecto en el que surge SAUTEE. En la Sección 4 se establecen las bases teóricas de la metodología con la cual opera el sistema, específicamente la selección de marcadores estilométricos, el cálculo de la similitud entre documentos y la técnica de visualización de los resultados. En la Sección 5 se describe a detalle el funcionamiento del sistema desde el punto de vista de la interfaz de usuario, seguido de un pequeño ejemplo de uso en la Sección 6. Finalmente, en la última sección se presentan conclusiones.

2 Recursos existentes

Con el fin de presentar los recursos existentes es importante tomar en cuenta que para hacer un análisis estilométrico de textos se siguen, en general, tres etapas: preprocesamiento, determinación de marcadores estilométricos y análisis estadístico. Este *pipeline* es muy común en tareas de procesamiento de textos y Juola et al. (2006) las describen para un sistema de atribución de autoría:

- **Preprocesamiento:** Son todas aquellas modificaciones que se hacen al texto antes de su procesamiento. Juola maneja esta fase bajo el nombre de canonización.
- **Determinación de marcadores estilométricos:** Es la especificación, lo que se va a medir en los textos. Por ejemplo, palabras, n-gramas de palabras, etc. Juola llama a esta fase selección del conjunto de eventos.
- **Selección del método de análisis:** Es la selección del método por el cual se hará el análisis estadístico para presentar conclusiones, resultados, gráficas, etc.

Bajo estos tres puntos se analizan 3 herramientas de análisis estilométricos, además de presentar sus ventajas y desventajas. Todas estas herramientas son gratuitas y están disponibles en la web para su descarga.

2.1 Signature

The Signature Stylometric System[1] es una aplicación de escritorio desarrollada por el profesor Peter Millican de la Universidad de Leeds.

2.1.1 Pipeline

Preprocesamiento. Signature hace un preprocesamiento mínimo de los textos. Uno de ellos es convertir todo el texto a mayúsculas. Además permite combinar varios textos en uno solo.

[1] Disponible en `http://www.philocomp.net/humanities/signature.htm`

Determinación de marcadores estilométricos. El programa realiza el conteo de unas características predeterminadas que son: distribución de longitud de palabras, oraciones y párrafos, frecuencia de letras y uso de signos de puntuación. El usuario puede adicionalmente ingresar una lista de palabras para incluirlas en el análisis. Fuera de esto no hay ninguna otra opción o parámetro relativo a esta etapa.

Selección de método de análisis. El método de análisis por defecto es la visualización de las frecuencias de los marcadores estilométricos en una gráfica del tipo histograma. Los corpus o documentos seleccionados aparecen cada uno con un color diferente. No hay otra manera de visualización, salvo la elección de ver la gráfica en 2D o 3D. El otro método de análisis es la prueba de la Chi cuadrada. Para esto se solicita que el usuario elija dos documentos (o un documento y un conjunto de documentos combinado). El programa determinará automáticamente si los rasgos presentes en los documentos permiten hacer la prueba (ya que la prueba requiere que los valores de las frecuencias rebasen cierto umbral). Si el análisis procede, el programa reportará el p-valor derivado de la prueba y su interpretación tradicional.

2.1.2 Ventajas

La aplicación ofrece la funcionalidad de "combinar archivos en corpus" por medio de la cual un conjunto de textos pueden ser combinados en uno solo. De esta manera se pueden combinar en un solo corpus todos los textos de un mismo autor, permitiendo comparar un documento individual (dubitado o texto de autoría desconocida) contra todos los demás textos de un determinado autor.

La gráfica resultante muestra de manera intuitiva las diferencias de cada marcador entre los distintos corpus. Además, Signature ofrece la prueba de la χ^2 para determinar la similitud entre dos textos de una manera cuantitativa.

2.1.3 Desventajas

Para visualizar los resultados, el programa crea una gráfica para palabras, otra para puntuación, otra para letras, etc. La gráfica generada no puede mostrar el acumulado de las diferencias de todos los marcadores al mismo tiempo, lo cual es una limitante. Por otro lado, el tipo de gráfica es impráctico para una lista de palabras de extensión considerable, puesto que termina haciéndose muy larga en el eje horizontal.

Además, el catálogo de marcadores estilométricos no es muy extenso y no tiene los más usuales, que son n-gramas de palabras y de caracteres (para $n > 1$).

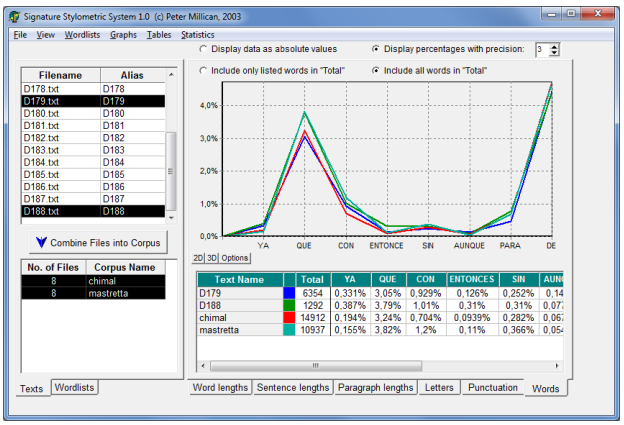

Figura 1: Interfaz de Signature.

2.2 JGAAP

JGAAP[2] es un proyecto desarrollado por Patrick Juola de la Universidad Duquesne (Juola, 2009), está diseñado para permitir a no-expertos en el área de aprendizaje de máquina un acercamiento a este tipo de técnicas, así como para facilitar la comparación entre la efectividad de varios métodos.

La interfaz gráfica (Figura 2) de este programa es muy organizada puesto que se presentan diferentes pestañas, cada una de las cuales indica las opciones para cada fase del pipeline.

2.2.1 Pipeline

Preprocesamiento. Ya que Juola maneja esta etapa como una parte esencial del análisis, la aplicación contiene una serie de opciones dedicadas únicamente a la misma. El usuario puede determinar exactamente qué tipo de procesamiento se hará al texto, como por ejemplo, eliminar caracteres especiales, eliminar signos de puntuación, entre otros.

Determinación de marcadores estilométricos. El programa cuenta con un extenso catálogo de diversos marcadores. El usuario puede elegir uno o más, y además especificar los parámetros de cada uno, de ser necesario. Por ejemplo, al elegir n-gramas se pedirá que se especifique el valor de n. De esta manera, el usuario puede pedir que se analicen bigramas o trigramas de palabras

[2]Disponible en https://github.com/evllabs/JGAAP

o bigramas o trigramas de etiquetas POS (*Part of Speech*), entre otros.

Selección del método de análisis. De manera similar a la determinación de marcadores, al usuario se le presenta una extensa lista de métodos de análisis, entre los que se encuentran, Análisis de Componentes Principales, Análisis Discriminante Lineal y Máquinas de Soporte de Vectores.

2.2.2 Ventajas

Contiene un extenso catálogo tanto de marcadores estilométricos como de métodos de análisis, los cuales son parametrizables.

2.2.3 Desventajas

La descripción de los marcadores estilométricos no es muy informativa.

La salida del programa es únicamente texto. No muestra ninguna gráfica ni diagrama, ni es posible exportar los datos a una hoja de cálculo.

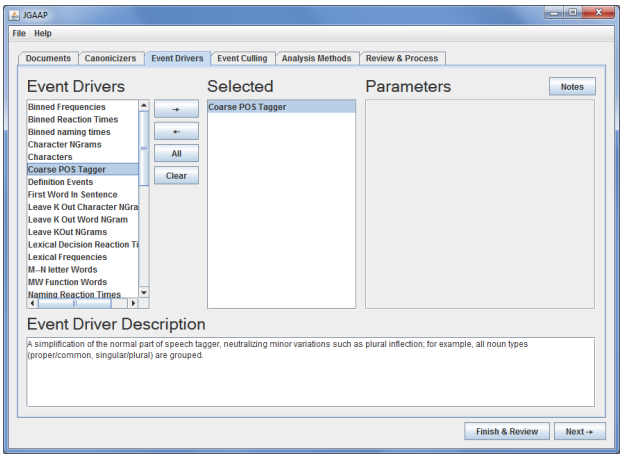

Figura 2: Interfaz de JGAAP.

2.3 Stylo

Stylo es un paquete diseñado para el entorno estadístico R (R Core Team, 2008). Ofrece una interfaz gráfica de usuario (Figura 3), de manera que no es necesario escribir un programa para poder usar sus funcionalidades.

2.3.1 Pipeline

Preprocesamiento. El procesamiento es mínimo: las únicas opciones que el usuario puede seleccionar son si desea que se preserven las mayúsculas (ya que por omisión este programa convierte todo el texto a minúsculas), y si desea que se eliminen los pronombres.

Determinación de marcadores estilométricos. Únicamente hay dos tipos de marcadores: palabras y caracteres. Sin embargo, hay varios parámetros manipulables por el usuario. Se puede elegir el tamaño de los n-gramas, así como el número de palabras que entrarán dentro del análisis. Es decir, se puede determinar que solo se usen las 100 palabras más frecuentes, o 50 o las que se deseen.

Selección del método de análisis. Se presenta al usuario las opciones para llevar a cabo el análisis estadístico, específicamente el tipo de análisis y el tipo de distancia.

2.3.2 Ventajas

Ofrece varias opciones de visualización de resultados, incluyendo escalamiento multidimensional, análisis de componentes principales y análisis de clusters. Asimismo, se puede hacer uso de varios métodos de clasificación como vecinos más cercanos, Bayes ingenuo, SVM, entre otros.

Además, gracias a que corre dentro del ambiente R, se puede hacer cualquier otro tipo de análisis con los datos generados, siempre que el usuario conozca el uso de este lenguaje.

2.3.3 Desventajas

Solo hace análisis con n-gramas de palabras y de caracteres, no hay ningún otro marcador estilométrico disponible.

La carga del corpus no es tan fácil como en las otras aplicaciones, pues se basa en preparar una estructura de carpetas determinada y seguir una convención para los nombres de los archivos.

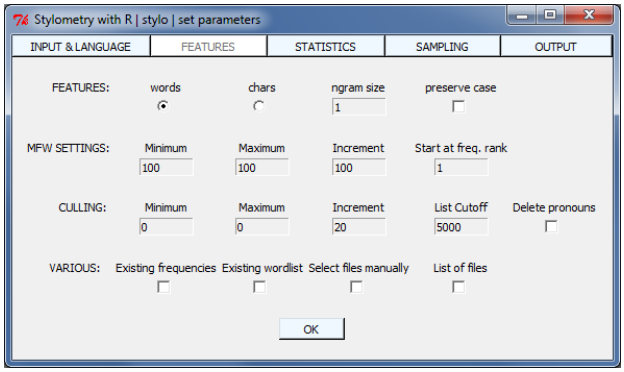

Figura 3: Interfaz de stylo.

3 Marco del proyecto

Una de las tareas del procesamiento automático de lenguaje natural que ha cobrado mayor importancia a últimas fechas es la detección de similitud textual. Esta labor responde a diversas necesidades, tales como la clasificación textual, la identificación de autoría, el análisis de reutilización de textos, la detección de paráfrasis y la detección de plagio. Con el fin de contribuir en el desarrollo de las metodologías existentes para la medición de similitud textual en documentos utilizando diferentes enfoques, tanto lingüísticos como estadísticos, se propuso la creación de un recurso lingüístico.

El objetivo fue generar una herramienta pública que sea de utilidad a los usuarios y fomente la colaboración. En este sentido, resultaba crucial la creación de un repositorio central de documentos, o gestor de corpus, en el que cualquier persona interesada pueda registrarse para poder cargar sus propios documentos. Este sistema tiene por nombre GECO: Sistema de Gestión de Corpus (Sierra et al., 2017). Los textos cargados pueden ser, opcionalmente, puestos a disposición de todos los demás usuarios de la plataforma, de manera que cada vez se puedan tener corpus más robustos.

El sistema presentado en este artículo, SAUTEE (Sistema Automático para Estudios Estilométricos)[3], es un sistema en web accesible desde cualquier computadora con conexión a Internet, el cual permite al usuario analizar la aparición de diversos marcadores estilométricos en un conjunto de documentos. El SAUTEE se alimenta de los documentos de los corpus creados a través de GECO, y es la primera de varias herramientas planeadas para sacar provecho de este repositorio de documentos. Cada herramienta tendrá un fin en específico, que en el caso del SAUTEE es el análisis estilométrico.

4 Fundamentos teóricos del sistema

El análisis realizado por el SAUTEE se basa en el cálculo de distancias entre los documentos. Para calcular la distancia que hay entre dos textos es preciso primeramente representar cada texto de una forma numérica, es decir, vectorizarlos. Una vez hecho esto y calculadas las distancias, se aplica un método de visualización de datos por medio del cual es posible apreciar la similitud entre cada documento. El SAUTEE realiza entonces las siguientes tres tareas:

- La vectorización de los textos, es decir, la extracción de marcadores estilométricos y su cuantificación.
- El cálculo de una distancia entre cada par de documentos.
- La generación de un conjunto de puntos en 2 dimensiones, por medio del cual se pueden visualizar en una gráfica las distancias calculadas.

4.1 Extracción de marcadores estilométricos y Vectorización

4.1.1 Marcadores de estilo

La estilometría se basa en identificar rasgos que puedan ser cuantificados y que sean característicos del estilo del autor. Bailey (1979) sugiere que estos rasgos deben ser:

- salientes;
- estructurales;
- frecuentes y fácilmente cuantificables;
- relativamente inmunes al control consciente.

Reciben en la literatura muchos nombres, uno de los más usuales es "marcadores de estilo". Normalmente este término hace referencia a la categoría general de la característica de la cual se está hablando. Por ejemplo, si se determina que la frecuencia de uso de ciertos signos de puntuación es una característica distintiva, entonces se dice que "signos de puntuación" es un marcador estilométrico y, dentro de esta categoría, existirán características específicas como lo son punto, coma, punto y coma, etcétera. En SAUTEE adoptamos el término "marcadores estilométricos" para referirnos a estas categorías.

A lo largo del tiempo se han identificado e investigado cientos de marcadores estilométricos — Rudman (1997) estima 1000 — sin embargo no se ha encontrado un conjunto que funcione para cualquier situación. Muchas veces el dominio en el que se lleva a cabo la investigación dictará el tipo de marcadores que se requieren. Por ejemplo, el tipo de marcadores estilométricos no será el mismo para analizar una novela que para analizar una publicación en una red social. Para hacer un recuento de los marcadores que son comúnmente utilizados en el área, se puede hacer una clasificación de acuerdo con diversos criterios. Por ejemplo, la presentada por Stamatatos (2009) en su revisión del tema, divide los marcadores en las siguientes categorías:

[3]Disponible en `http://www.corpus.unam.mx/saute`

- **Léxicos.** Aquí se encuentran todas aquellas medidas a nivel palabra, por ejemplo n-gramas de palabras, longitud de palabras y oraciones, riqueza de vocabulario, etc.
- **Caracter.** Son aquellas medidas a nivel caracter, por ejemplo n-gramas de caracteres y conteos de caracteres específicos (por ejemplo dígitos).
- **Sintácticos.** Primordialmente se incluyen aquí las etiquetas POS (Part of speech) que identifican a cada palabra con su parte de la oración. Estas etiquetas pueden contener tanta información como sea necesario, por ejemplo, para el caso de los verbos: persona gramatical, número, tiempo verbal, etc. La estructura sintáctica de una frase puede ser representada por n-gramas de etiquetas de parte de la oración, aunque también pueden usarse soluciones más complejas como parsers o chunkers.
- **Semánticos.** Statamatos cita tres principales: dependencias semánticas, análisis de sinónimos e hiperónimos, y "características funcionales". Estas últimas asignan un rol a las palabras dentro del discurso, por ejemplo "elaboración", "clarificación", etc.
- **Específicos a la aplicación.** Son aquellas características que dependen del dominio en el que se está llevando a cabo el experimento y que pueden ser estructurales, específicas del contenido o específicas del lenguaje. Por ejemplo, al analizar textos de internet podemos extraer características como nombres de usuario, etiquetas HTML, entre otras.

Para lograr extraer cada tipo de característica se necesitan herramientas diferentes. Por ejemplo, para poder hacer análisis usando información de etiquetas POS, se necesita una herramienta capaz de hacer este etiquetado. En el caso de SAUTEE, esta herramienta es Freeling (Padró & Stanilovsky, 2012). Freeling es una suite de análisis del lenguaje desarrollada en la Universidad Politécnica de Cataluña bajo la dirección de Lluís Padró.

4.1.2 Catálogo de Marcadores del SAUTEE

A continuación se describen los marcadores estilométricos con los que cuenta SAUTEE al momento de escribir este artículo. Todos ellos se basan en el conteo de las apariciones de ciertas características en el texto. Además, ya que cada texto tiene diferente longitud, se lleva a cabo una normalización de manera que las frecuencias utilizadas son relativas.

Signos de puntuación. Se contabilizan los signos de puntuación del texto con base en el etiquetado de Freeling (es decir, se toma como signo de puntuación todo lo que Freeling etiqueta como tal). Cada frecuencia se divide entre el número total de signos de puntuación en el texto.

Distribución de longitud de oraciones y palabras. Se contabilizan las frecuencias de aparición de las siguientes categorías de palabras: palabras de 1 letra, palabras de 2 letras, palabras de 3 letras, sucesivamente hasta 20 letras. Cada frecuencia se divide entre el número total de palabras en el texto. Respecto a la longitud de las oraciones se tienen las categorías: menos de 10 palabras, de 11 a 20, de 21 a 30, de 31 a 40, de 41 a 50, y más de 51. Cada frecuencia se divide entre el número total de oraciones en el texto.

Categoría gramatical al inicio de la oración. A partir de las etiquetas POS generadas por Freeling, se contabiliza el número de veces que cada categoría gramatical aparece al inicio de una oración. Por ejemplo, cuántas veces un verbo inicia la oración, cuántas veces un sustantivo, y así sucesivamente. Cada frecuencia es dividida entre el número total de palabras al inicio de la oración (o lo que es lo mismo, entre el número total de oraciones del texto).

Categoría gramatical al final de la oración. Lo mismo que la anterior, pero considerando las palabras al final de la oración.

Unigramas de palabras funcionales. Se contabiliza la aparición de las palabras funcionales (de acuerdo con la lista de palabras funcionales cargada actualmente en el sistema). La frecuencia de cada palabra se divide entre el total de apariciones de palabras funcionales contabilizadas en el texto.

Bigramas de palabras funcionales. Lo mismo que la anterior, pero considerando bigramas, es decir todas aquellas apariciones de dos palabras funcionales seguidas. Por ejemplo, tomando como palabras funcionales los artículos y las conjunciones, en el segmento "el niño y la niña" se contabilizaría como bigrama de palabras funcionales "y la". La frecuencia de cada bigrama es dividida entre el total de bigramas de palabras funcionales contabilizadas en el texto.

Trigramas de palabras funcionales. Lo mismo que la anterior pero tomando en cuenta

las apariciones de tres palabras funcionales seguidas. La frecuencia de cada trigrama es dividida entre el total de trigramas de palabras funcionales contabilizadas en el texto.

Bigramas de palabras funcionales con hasta 2 huecos. En este caso se contabilizan las apariciones de dos palabras funcionales que no están contiguas, sino que están separadas a lo más por otras dos palabras. Por ejemplo, en la frase "el niño y la niña", se contabilizaría el bigrama "el y", cuyos elementos están a una palabra de separación (en este caso, "niño"). Cada frecuencia se divide entre el total de bigramas de palabras funcionales con hasta 2 huecos contabilizadas en el texto.

Trigramas de palabras funcionales con hasta 2 huecos. Lo mismo que la anterior pero considerando apariciones de tres palabras funcionales. No necesariamente tiene que haber el mismo número de huecos entre la primera y la segunda palabra funcional que entre la segunda y la tercera. Por ejemplo, la primera y la segunda palabra funcional pueden estar a una separación de una palabra, y la segunda y la tercera a una distancia de dos. Cada frecuencia se divide entre el total de trigramas de palabras funcionales con hasta 2 huecos contabilizadas en el texto.

Unigramas de etiquetas POS. Se contabiliza la aparición de las etiquetas POS tal como Freeling las genera. La frecuencia de cada etiqueta se divide entre el número de palabras en el texto.

Bigramas de etiquetas POS. Se contabilizan las apariciones de dos etiquetas POS contiguas. La frecuencia de cada bigrama se divide entre el total de bigramas de etiquetas POS contabilizadas en el texto.

Trigramas de etiquetas POS. Lo mismo que la anterior pero considerando tres etiquetas contiguas. La frecuencia de cada trigrama se divide entre el total de trigramas de etiquetas POS contabilizadas en el texto.

Unigramas de etiquetas POS no fino. Lo mismo que unigramas de etiquetas POS pero en vez de contabilizar la frecuencia de las etiquetas tal cual las genera Freeling, se toma en cuenta únicamente el primer caracter de la etiqueta que corresponde a la categoría gramatical más general. Por ejemplo, la etiqueta vmii3s0 (verbo principal indicativo imperfecto tercera persona de singular) y la etiqueta vsip3p0 (verbo semiauxiliar imperfecto tercera persona del plural) se agrupan bajo una misma etiqueta "v", verbo. La frecuencia de cada una de estas etiquetas simplificadas se divide entre el número total de palabras.

Bigramas de etiquetas POS no fino. Lo mismo que la anterior pero contabilizando las apariciones de dos etiquetas seguidas. La frecuencia de cada bigrama se divide entre el total de bigramas de etiquetas POS no fino contabilizadas en el texto.

Trigramas de etiquetas POS no fino. Lo mismo que la anterior pero contabilizando las apariciones de tres etiquetas seguidas. La frecuencia de cada trigrama se divide entre el total de trigramas de etiquetas POS no fino contabilizadas en el texto.

Bigramas de caracteres. Se contabilizan las apariciones de dos caracteres seguidos. Los espacios se consideran caracteres. Por ejemplo en el segmento "el niño", los bigramas de caracteres son "el", "l_", "_n", "ni", "iñ", "ño" (el guión bajo representa un espacio). La frecuencia de cada bigrama se divide entre el total de bigramas de caracteres contabilizados en el texto.

Trigramas de caracteres. Lo mismo que la anterior pero considerando tres caracteres seguidos. La frecuencia de cada trigrama se divide entre el total de trigramas de caracteres contabilizados en el texto.

4.1.3 Vectorización del texto

Una vez seleccionados los marcadores se crea un vector por cada documento. A continuación se presenta un ejemplo de la creación de estos vectores. Sea el texto (1),

> *La cantante de ópera deleitó al público en la función de anoche.*

El preprocesamiento hecho por Freeling obtiene el lema y la etiqueta POS de cada palabra del texto, como se puede ver en el cuadro 1.

De acuerdo a los marcadores estilométricos elegidos por el usuario, se hacen los respectivos conteos de aparición en el texto. Por ejemplo, sea el marcador elegido "Unigramas de etiquetas POS no fino", las características obtenidas

Palabra	Lema	POS
La	el	DA0FS0
cantante	cantante	NCCS000
de	de	SPS00
ópera	ópera	NCFS000
deleitó	deleitar	VMIS3S0
a	a	SPS00
el	el	DA0MS0
público	público	NCMS000
en	en	SPS00
la	el	DA0FS0
función	función	NCFS000
de	de	SPS00
anoche	anoche	RG

Cuadro 1: Análisis de Freeling para el texto (1).

usando este marcador y sus respectivos valores se pueden ver en el cuadro 2.

Característica	Frecuencia relativa
D	23.08 %
N	30.77 %
S	30.77 %
V	7.69 %
R	7.69 %

Cuadro 2: Frecuencias relativas para el texto (1).

Sea el texto (2),

> *El mesero trajo la sopa fría y una hora tarde.*

Sus características y frecuencias se muestran en el cuadro 3. Aquí debemos notar que hay dos características que no están presentes en el texto anterior: la conjunción (C) y el adjetivo (A), y que le falta una característica que sí está presente en el primero, la preposición (S). Por lo tanto para que ambos vectores sean susceptibles de ser comparados, deben modificarse para que contengan el mismo número de características. Las características que el otro texto no comparte se llenan con ceros. En el cuadro 4 se puede ver cómo quedan ambos vectores después de hacer las modificaciones necesarias. La siguiente etapa en el análisis es medir la similitud entre los textos mediante la aplicación de una función de distancia a estos vectores.

4.2 Medidas de distancia

Una medida de distancia o métrica es un valor que se define para cada par de elementos de un conjunto, en este caso el conjunto de textos.

Característica	Frecuencia relativa
D	30.00 %
N	30.00 %
V	10.00 %
A	10.00 %
C	10.00 %

Cuadro 3: Frecuencias relativas para el texto (2).

	Texto 1	Texto 2
D	23.08	30
N	30.77	30
S	30.77	0
V	7.69	10
R	7.69	10
A	0	10
C	0	10

Cuadro 4: Vectores para el texto (1) y el texto (2).

Busca cuantificar la disimilitud entre cada par de elementos de manera que la medida tenga una magnitud mayor para elementos no semejantes. Es una función que toma como entrada dos elementos y da como resultado un único número. Si el resultado de aplicar la función a los vectores de dos textos es un número pequeño significa que éstos poseen un estilo similar. Desde luego el significado de pequeño variará de acuerdo al tipo de medida seleccionada, y dependerá del valor de los demás elementos del corpus, es decir, es un valor relativo. A continuación se presentan las distancias con que el sistema cuenta actualmente.

4.2.1 Distancia Euclidiana

Para todo par de puntos a y b, la distancia euclidiana representa el camino más corto entre ellos, es decir una línea recta. Es la distancia que puede resultar más intuitiva puesto que para el caso de 2 y 3 dimensiones, es equivalente a nuestra idea de distancia en el mundo real; sin embargo, puede ser generalizable a cualquier número de dimensiones. En este caso, las dimensiones corresponden a cada una de las características contabilizadas. Matemáticamente, la distancia euclidiana es igual a la raíz cuadrada de la suma del cuadrado de las diferencias de cada dimensión y se expresa por medio de la siguiente fórmula:

$$\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

En donde estamos hablando de que hay n características en cada documento, y que X y Y representan los vectores de los dos documentos, es decir, que X_i representa el valor de la i-ésima característica en el primer documento y Y_i representa el valor de esa misma característica en el segundo documento.

4.2.2 Distancia Manhattan

La distancia Manhattan es también llamada distancia de taxista, haciendo referencia a la distancia que un vehículo tendría que recorrer para llegar de un punto a otro en una cuadrícula. Matemáticamente es igual a la suma de las diferencias absolutas entre cada dimensión, como lo muestra la siguiente ecuación:

$$\sum_{i=1}^{n} |X_i - Y_i|$$

4.2.3 Delta de Burrows

La Delta de Burrows es un método específicamente desarrollado para medir la diferencia estilística entre un conjunto de documentos. Originalmente publicado en 2002, se ha convertido en un referente de los estudios de autoría. Se basa en contar la frecuencia de un conjunto de palabras en un texto y calcular el z-score de cada una, es decir, su número de desviaciones estándar sobre la media. La Delta, tal y como fue definida por Burrows, es el promedio de las diferencias absolutas entre los z-scores de un conjunto de palabras en un grupo de textos y los z-scores del mismo conjunto de palabras en el texto objetivo (Burrows, 2002).

Stein & Argamon (2006) demuestram, que matemáticamente esta definición es equivalente a una distancia Manhattan ponderada, en donde el peso en cada dimensión corresponde a la desviación estándar de esa dimensión. Específicamente, es igual a:

$$\sum_{i=1}^{n} \frac{|X_i - Y_i|}{\sigma_i}$$

En donde X son las palabras correspondientes al primer texto y Y las palabras correspondientes al segundo texto, n es el total de palabras y σ es la desviación estándar de la frecuencia de cada palabra. SAUTEE hace el cálculo de la Delta usando esta formula.

4.2.4 Distancia Canberra

La distancia Canberra es otro ejemplo de una distancia Manhattan ponderada. En este caso, la diferencia absoluta entre las variables es dividida entre la suma de los valores absolutos de las mismas. Es igual a:

$$\sum_{i=1}^{n} \frac{|X_i - Y_i|}{|X_i| + |Y_i|}$$

Esta distancia tiene la bondad de ser sensible a valores cercanos a cero, por lo cual es útil si el conjunto de datos contiene, tanto valores pequeños, como valores muy grandes.

4.3 Escalamiento multidimensional

El escalamiento multidimensional (MDS por sus siglas en inglés) es una técnica en el área de visualización de datos que permite apreciar las distancias existentes entre un conjunto de objetos. Su objetivo es asignar a cada punto de un conjunto de datos de n-dimensiones, una coordenada en un espacio de menor dimensión (comúnmente 2), de tal modo que las distancias entre los puntos en este nuevo espacio se mantengan lo más parecidas posible a las distancias originales. De esta manera los puntos pueden ser graficados en un plano cartesiano donde se puede ver la relación que existe entre cada uno respecto a los demás. En este caso cada punto representará un documento y su cercanía o lejanía con otros documentos indicará qué tan similar es el estilo de ambos.

El MDS recibe como entrada una matriz de disimilaridad cuyos elementos d_{ij} representan la distancia que hay entre el objeto i y j. El objetivo en concreto del escalamiento multidimensional es encontrar un conjunto de vectores $x_1, ..., x_n$, $x \in \mathbb{R}^N$ tal que $D(x_i, x_j) \approx d_{ij}$ donde N es la nueva dimensionalidad deseada y D es normalmente la distancia euclidiana. De esta manera, si $N = 2$, los vectores resultantes serán coordenadas de dos dimensiones que pueden ser graficadas sin problema en un plano cartesiano o gráfico de dispersión. Por ejemplo, si los datos de entrada fueran las distancias existentes entre un conjunto de ciudades europeas, tras llevar a cabo el MDS se obtendría un conjunto de puntos $\mathbb{R}^2$ cuya gráfica en el plano cartesiano se asemejaría a un mapa de Europa.

Una de las ventajas de hacer un análisis con MDS es que se puede visualizar el efecto que tienen simultáneamente todos los marcadores estilométricos elegidos. Por lo tanto, se pueden hacer

experimentos utilizando diversas combinaciones de estos.

Es importante poder visualizar este efecto simultáneo de un conjunto determinado de marcadores puesto que los marcadores que son discriminativos para un autor no necesariamente lo serán para otro, y por lo tanto resulta ilustrativo hacer varias pruebas con conjuntos de marcadores diferentes.

5 Funcionamiento del sistema

En la figura 4 se puede observar la interfaz principal del SAUTEE. Tiene un diseño basado en pestañas, cada una de las cuales representa un paso en la secuencia del proceso completo (similar al JGAAP). SAUTEE funciona de la siguiente manera: primeramente se solicita al usuario que seleccione los documentos que desea analizar. Después han de elegirse los marcadores estilométricos que serán tomados en cuenta en el análisis. Finalmente es necesario indicar qué tipo de distancia intertextual será calculada entre cada par de documentos. El tipo de distancia elegida puede depender del tipo de marcador estilométrico que se esté analizando (López-Escobedo et al., 2016). Tras llevar a cabo los pasos anteriores, el sistema genera una gráfica que muestra visualmente la distancia entre cada par de documentos, y da al usuario la opción de descargar las estadísticas en un formato de hoja de cálculo.

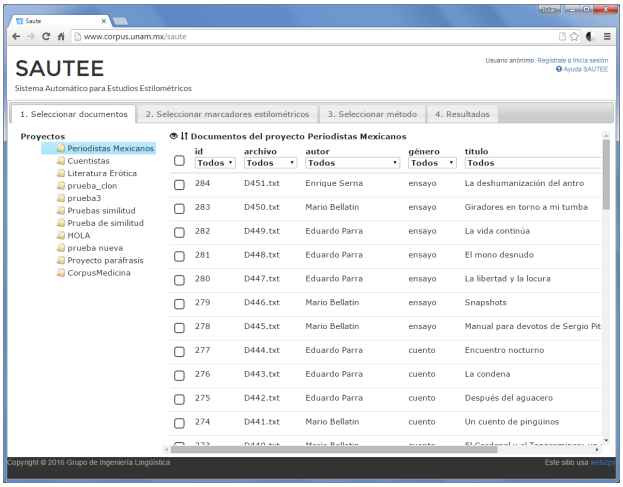

Figura 4: Interfaz del SAUTEE.

5.1 Pipeline general

5.1.1 Preprocesamiento

El preprocesamiento no es llevado a cabo directamente por el SAUTEE. Como se ha mencionado, los textos con los que opera el SAUTEE deben ser cargados primero en la plataforma GECO. Una vez cargados ahí automáticamente pasan por un preprocesamiento que consiste en las siguientes tareas:

- Convierte el archivo a texto plano (en el caso de documentos subidos como PDF o DOC).
- Codifica el texto en UTF-8 (si tenía el texto una codificación distinta).
- Entrega los textos a Freeling para su lematización y etiquetado de partes de la oración (Part of Speech o POS).

Gracias a este preprocesamiento, el SAUTEE trabaja con textos que ya están separados por oraciones y palabras, y en los cuales cada palabra está anotada con su lema y su etiqueta POS. De estos textos preprocesados se extraen las características con las cuales se hará el análisis.

5.1.2 Determinación de características

El SAUTEE solicita al usuario que elija uno o varios marcadores estilométricos de entre los disponibles en el catálogo del sistema y que se describen en el apartado 4.1.2. Esta selección determinará los marcadors con los que se construirá el vector del documento. Si se selecciona más de un marcador, el vector contendrá características de cada uno de ellos.

5.1.3 Selección del método de análisis

El SAUTEE calcula una medida de distancia entre cada par de vectores generados en la etapa anterior. La medida de distancia puede ser especificada por el usuario. A partir de las distancias generadas se realiza el escalamiento multidimensional para producir una gráfica en 2 dimensiones que es presentada al usuario para su análisis.

5.2 Operación a detalle

5.2.1 Selección de documentos

En esta pantalla el sistema muestra un listado de los corpus disponibles. Al seleccionar un corpus el sistema muestra un listado de los documentos individuales que lo conforman. Estos documentos pueden, además, venir acompañados

de una serie de metadatos, como lo son, autor, género literario, entre otros. De esta lista de documentos se tienen que elegir por lo menos dos.

5.2.2 Selección de marcadores estilométricos

En este apartado se muestra un listado de los marcadores estilométricos actualmente disponibles para el análisis y que fueron enumerados en el apartado 4.1.2 de este documento. Como se mencionó en el marco del proyecto, el sistema está pensado para recibir actualizaciones frecuentemente de manera que esta lista de maracadores estilométricos se vea aumentada respondiendo a sugerencias de los usuarios o a avances en el estado del arte del área. Esta pantalla muestra también una pequeña descripción de cada marcador estilométrico, especificando exactamente la manera en que es calculado para un texto dado. El usuario debe seleccionar por lo menos un marcador, pero pueden seleccionarse cualquier número de ellos, incluso todos. El vector generado para calcular las distancias contendrá características de todos los marcadores seleccionados.

5.2.3 Selección de método

La selección de método hace referencia a la forma en que será calculada la distancia entre cada par de documentos. Se incluye en el sistema la explicación de cada método y la bibliografía correspondiente, en su caso. Una vez elegido el método todo está listo para comenzar el análisis. En esta misma pantalla se encuentra un botón que dispara el proceso, el cual puede durar desde unos pocos segundos a algunos minutos dependiendo del volumen de los textos, así como de la cantidad de marcadores estiolométricos elegidos. Una vez que el proceso concluye se muestra la sección de resultados.

5.2.4 Resultados

Esta es la última pantalla, en donde el usuario visualiza el resultado del análisis. Tras el proceso de generación de distancias intertextuales, el método de escalamiento multidimensional asigna a cada documento una coordenada en un espacio de dos dimensiones. Todos estos puntos son mostrados en un diagrama de dispersión de manera que la distancia aparente entre los puntos es proporcional a la distancia intertextual realmente calculada entre los documentos. De esta manera, el usuario puede visualizar la similitud estilística entre todos los documentos inmediatamente. Adicionalmente, los datos numéricos resultantes del proceso se preparan en dos archivos en formato CSV (visualizable en cualquier programa de hoja de cálculo) que el usuario puede descargar para análisis subsecuentes. Uno de estos archivos contiene la frecuencia de uso de cada característica perteneciente a los marcadores estilométricos seleccionados (es decir, los vectores utilizados para calcular las distancias). El otro archivo contiene las distancias intertextuales calculadas entre cada par de documentos (las distancias entre los vectores).

6 Caso de uso

Veamos un ejemplo de uso. Uno de los corpus disponibles por defecto es un corpus llamado Periodistas Mexicanos. Contiene artículos, ensayos y cuentos de 6 escritores mexicanos o residentes en México desde temprana edad: Alberto Chimal, Ángeles Mastretta, Enrique Serna, José de la Colina, Eduardo Parra y Mario Bellatín. Tiene 9 documentos de cada uno de estos autores, dando un total de 54 documentos. En este ejemplo se hará la comparación entre Ángeles Mastretta y José de la Colina.

Se empieza por seleccionar el corpus Periodistas Mexicanos de la lista de proyectos del lado izquierdo. Una vez seleccionado aparecerán del lado derecho los documentos que conforman este corpus. En la columna de autor abrimos el filtro y se selecciona Ángeles Mastretta. Para seleccionar todos los documentos de la autora se selecciona la casilla que se encuentra en el encabezado de la tabla en la primera columna. Se repite el mismo procedimiento pero ahora seleccionando en el filtro a José de la Colina.

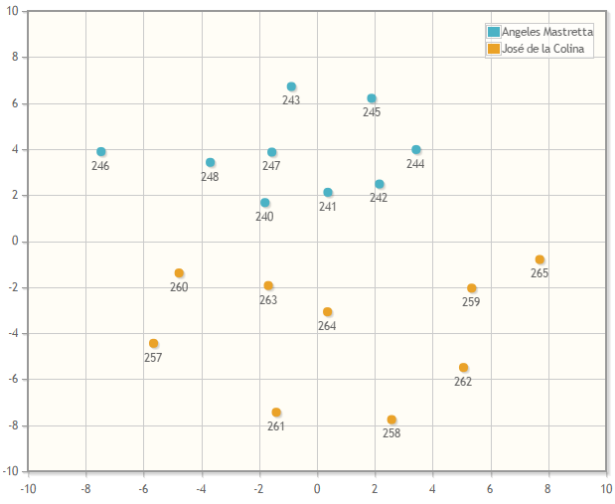

Figura 5: Análisis con unigramas POS. Coloreado por autor.

Una vez seleccionados los 18 documentos se procede a la siguiente pestaña para hacer la selección de marcadores estiolmétricos. Para el primer experimento se selecciona un único marcador: unigramas de etiquetas POS. En la siguiente pestaña se selecciona distancia euclidiana. Una vez terminado el procesamiento se obtiene la gráfica mostrada en la figura 5 (seleccionando la opción de colorear por autor). En la gráfica podemos observar que los dos grupos de textos se separan visiblemente. De hecho, todos los textos de Ángeles Mastretta quedan por arriba del eje x y todos los textos de José de la Colina por debajo. Se puede concluir que los unigramas de etiquetas POS es un buen marcador para diferenciar estos dos autores.

Para un segundo experimento, se seleccionan esta vez unigramas, bigramas y trigramas POS simultáneamente y se vuelve a seleccionar distancia euclidiana. La gráfica resultante se muestra en la figura 6 (esta vez se selecciona la opción de colorear por género literario). Se puede observar que en esta gráfica los dos grupos de textos que se forman claramente son uno conformado por cuentos, y otro conformado por los artículos y ensayos. Se puede concluir que la combinación de unigramas, bigramas y trigramas de etiquetas POS no es un buen marcador en este caso para diferenciar el autor de los textos, sin embargo es útil para diferenciar el estilo del género literario cuento.

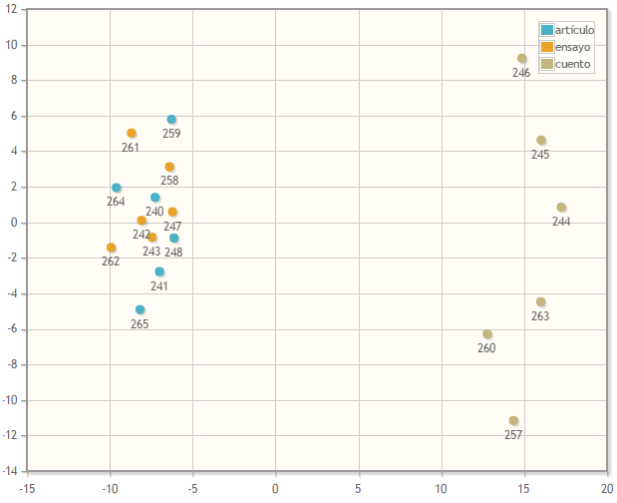

Figura 6: Análisis con unigramas, bigramas y trigramas POS. Coloreado por género literario.

7 Conclusiones

En este artículo se ha presentado un sistema que lleva a cabo extracción de características estilísticas en un corpus especificado por el usuario, que es cargado a un repositorio central en la nube por medio de una herramienta adicional llamada GECO. Es un sistema web, lo cual lo distingue de otras aplicaciones existentes, y permite que haya una mejor administración de los corpus. Toma en consideración un catálogo de marcadores estilométricos que no necesariamente es estático sino que puede ser ampliado por medio de parametrización. De cualquier modo, está pensado para añadir nuevos marcadores conforme se vaya necesitando y se propongan nuevos. Por otro lado, la presentación de los resultados en el sistema es por medio de una técnica de escalamiento multidimensional, lo cual permite intuitivamente apreciar la similitud y disimilitud entre todos los documentos del corpus, y usando una combinación de cualquier número de marcadores estilométricos.

Agradecimentos

Se agradece el apoyo recibido por el Consejo Nacional de Ciencia y Tecnología, proyecto 2016-01-2225 y a DGAPA-PAPIIT proyecto IA401517 y proyecto IA401419.

Referencias

Bailey, Richard W. 1979. Authorship attribution in a forensic setting. En *Advances in computer-aided literary and linguistic research*, 1–20. AMLC.

Burrows, John. 2002. Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17(3). 267–287. doi 10.1093/llc/17.3.267.

Coulthard, Malcolm. 1994. On the use of corpora in the analysis of forensic texts. *International Journal of Speech Language and the Law* 1(1). 27–43. doi 10.1558/ijsll.v1i1.27.

Diederich, Joachim, Jörg Kindermann, Edda Leopold & Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied intelligence* 19(1-2). 109–123. doi 10.1023/A:1023824908771.

Holmes, David I. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing* 13(3). 111–117. doi 10.1093/llc/13.3.111.

Juola, Patrick. 2009. JGAAP: a system for comparative evaluation of authorship attribution. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1(1). doi 10.6082/M1N29V4Z.

Juola, Patrick. 2015. The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities* 30(suppl 1). i100–i113. 10.1093/llc/fqv040.

Juola, Patrick, John Sofko & Patrick Brennan. 2006. A prototype for authorship attribution studies. *Literary and Linguistic Computing* 21(2). 169–178. 10.1093/llc/fql019.

López-Escobedo, Fernanda, Julián Solorzano-Soto & Gerardo Sierra Martínez. 2016. Analysis of intertextual distances using multidimensional scaling in the context of authorship attribution. *Journal of Quantitative Linguistics* 23(2). 154–176. 10.1080/09296174.2016.1142324.

Mendenhall, Thomas Corwin. 1887. The characteristic curves of composition. *Science* 9(214). 237–249.

Mosteller, Frederick & David Wallace. 1964. *Inference and disputed authorship: The federalist.* Addison-Wesley.

Nieto, Victoria Guillén, Chelo Vargas Sierra, María Pardiño Juan, Patricio Martínez Barco & Armando Suárez Cueto. 2008. Exploring state-of-the-art software for forensic authorship identification. *International Journal of English Studies* 8(1). 1–28.

Padró, Lluís & Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. En *Language Resources and Evaluation Conference (LREC)*, 2473–2479.

R Core Team. 2008. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Rudman, Joseph. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities* 31(4). 351–365. 10.1023/A:1001018624850.

Sierra, Gerardo, Julián Solórzano Soto & Arturo Curiel Díaz. 2017. GECO, un gestor de corpus colaborativo basado en web. *Linguamática* 9(2). 57–72. 10.21814/lm.9.2.256.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3). 538–556. 10.1002/asi.v60:3.

Stein, Sterling & Shlomo Argamon. 2006. A mathematical explanation of burrows's delta. En *Digital Humanities Conference*, 207–209.

Svartvik, Jan. 1968. *The evans statements: A case for forensic linguistics.* University of Gothenburg.

Tweedie, Fiona J., Sameer Singh & David I. Holmes. 1996. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities* 30(1). 1–10. 10.1007/BF00054024.

http://www.linguamatica.com/

linguaMATICA