Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 15, Número 2 (2023)

ISSN: 1647-0818

Volume 15, Número 2 – 2023

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Revisão

A comissão científica da **Linguamática** pode ser consultada na página Web da revista, em https://linguamatica.com/index.php/linguamatica/about/editorialTeam.

Para esta edição, colaboraram os seguintes investigadores:

- **Alberto Álvarez Lugrís**, Universidade de Vigo
- **António Teixeira**, Universidade de Aveiro
- **Cláudia Freitas**, ICMC, Universidade de São Paulo
- **David Soares Batista**
- **Diana Santos**, Linguateca / Universidade de Oslo
- **Gerardo Sierra**, Universidad Nacional Autónoma de México
- **Hugo Gonçalo Oliveira**, Universidade de Coimbra
- **Irene Castellón Masalles**, Universitat de Barcelona
- **Margarita Alonso-Ramos**, Universidade da Coruña
- **Mário Rodrigues**, Universidade de Aveiro
- **Renata Vieira**, Universidade de Évora
- **Ricardo Rodrigues**, Universidade de Coimbra
- **Sandra Collovini**, PUCRS
- **Sandra Maria Aluisio**, ICMC, Universidade de São Paulo
- **Sara Carvalho**, NOVA CLUNL e Universidade de Aveiro

Artigos de Investigação

Proposta recebida em Junho 2023 e aceite para publicação em Agosto 2023.

Desenvolvimento e avaliação de um modelo NER no domínio da análise cultural e do turismo

Development and evaluation of a NER model in the domain of cultural analysis and tourism

Susana Sotelo Docío
Universidade de Santiago de Compostela

Pablo Gamallo
Universidade de Santiago de Compostela

Álvaro Iriarte
Universidade do Minho

Resumo

O Reconhecimento de Entidades Mencionadas (NER) é uma tarefa essencial de extracção de informação em que as entidades de um texto são identificadas e classificadas. Um dos principais desafios enfrentados pelos sistemas NER é a dificuldade de generalização do aprendido para outros tipos de corpora diferentes dos utilizados durante o treino. Este problema é acentuado pelo facto de a maioria dos corpora de treino utilizados serem de natureza jornalística e, portanto, precisarem de ser adaptados a outros géneros e domínios. Neste artigo, utilizamos um corpus espanhol composto por entrevistas a visitantes da cidade de Santiago de Compostela e anotado com entidades mencionadas, para a avaliação e treino de sistemas NER adaptados ao domínio da cultura e do turismo. Apresentamos uma comparação das diferentes abordagens aplicadas, desde algoritmos clássicos de aprendizagem automática ao afinamento de vários modelos de *Transformers*. Os resultados obtidos superam significativamente o *baseline*, representado aqui pelos *toolkits* Stanza, spaCy e Flair, embora os testes preliminares com entidades não observadas durante o treino sugiram a necessidade de avaliações adicionais da sua capacidade de generalização e o uso de um método de segmentação adversarial no corpus.

Palavras chave

reconhecimento de entidades mencionadas, aprendizagem automática, redes neuronais, transformers, avaliação

Abstract

Named Entity Recognition (NER) is an essential task in information extraction where entities in a text are identified and classified. One of the primary challenges addressed by NER systems is the difficulty of generalizing what was learned to different types of corpora beyond the training data. This problem is magnified by the fact that most of the training corpora used are journalistic and therefore need to be adapted to other genres and domains. In this paper, we use a Spanish corpus consisting of interviews with visitors to the city of Santiago de Compostela and annotated with named entities, to evaluate and train NER systems tailored to the domain of cultural analysis and tourism. We provide a comprehensive comparison of various approaches employed, ranging from classical machine learning algorithms to fine-tuning Transformer models. The results significantly outperform the baseline, represented here by the toolkits Stanza, spaCy and Flair, although initial tests with unseen entities during training highlight the need for additional evaluations regarding their generalization capability and the utilization of adversarial splits for the corpus.

Keywords

named-entity recognition, machine learning, neural networks, transformers, evaluation

1. Introdução

O Reconhecimento de Entidades Mencionadas (NER, *Named-Entity Recognition*) é uma tarefa de extração de informação que consiste em identificar e classificar entidades tais como lugares, pessoas ou organizações. O NER é de grande utilidade para diversas tarefas em processamento de linguagem natural: por exemplo, em análise de sentimentos permite identificar a entidade sobre a que se emite uma opinião (Kanev et al., 2022; Barachi et al., 2022), e em tradução automática pode servir para selecionar aquelas entidades que não devem ser traduzidas (Vu et al., 2020; Lee et al., 2022). Tem sido aplicado a múltiplos domínios específicos, como saúde, segurança ou a extração de nomes em textos jornalísticos. No

DOI: 10.21814/lm.15.2.405

entanto, os modelos NER têm dificuldade em generalizar o que aprendem e o seu desempenho degrada-se ao serem aplicados a domínios muito diferentes daqueles para os quais foram treinados (Augenstein et al., 2017, pp. 70–73). Este problema é acentuado pelo facto de a maioria dos corpora anotados existentes serem de natureza jornalística e, portanto, precisarem de ser adaptados a outros géneros e domínios.

Neste trabalho utilizaremos um corpus em espanhol composto por entrevistas a visitantes da cidade de Santiago de Compostela e anotado com entidades mencionadas, para a avaliação e treino de sistemas NER adaptados ao domínio da cultura e do turismo. Mostraremos uma comparação de diferentes abordagens, que vão desde algoritmos clássicos de *machine learning* ao *fine-tuning* de vários modelos de *Transformers*. Os resultados das experiências realizadas melhoram amplamente o ponto de partida (*baseline*), representado aqui pelos modelos NER generalistas integrados nos *toolkits* Stanza (Qi et al., 2020), spaCy[1] e Flair (Akbik et al., 2019).

Tanto os modelos desenvolvidos quanto o corpus têm como objetivo final a extração automática de informação para o estudo das narrativas culturais relacionadas com a cidade de Santiago de Compostela e o Caminho de Santiago. Essa abordagem está fundamentada na conceção de cultura como o conjunto de mecanismos através dos quais os indivíduos e as comunidades organizam as suas vidas e as suas visões e, portanto, como um fenómeno social suscetível de análise, do que deriva a ideia de comparar os corpora narrativos com a realidade social (Torres Feijó, 2019).

Os principais contributos deste trabalho são:

- uma avaliação de diversas abordagens para NER aplicadas a uma combinação pouco explorada de domínio, registo e língua.
- uma análise da generalização dos modelos produzidos para obter uma ideia mais precisa do seu desempenho.
- a comparação da distribuição de entidades em diferentes conjuntos de dados, incluindo o corpus de trabalho, e a discussão sobre como isso pode impactar o desempenho de um modelo NER.

O artigo está estruturado da seguinte forma: em primeiro lugar, descrevemos o corpus utilizado e os resultados da avaliação de várias ferramentas de processamento de linguagem natural para determinar o seu grau de adaptação ao domínio específico. Seguidamente, apresentamos as diferentes abordagens consideradas, juntamente com os modelos treinados em cada uma delas, bem como os resultados da avaliação desses modelos. Por fim, descrevemos brevemente as linhas de trabalho futuro, incluindo uma análise preliminar de uma nova abordagem baseada em aprendizagem em contexto (engenharia de *prompting*).

1.1. Trabalho relacionado

A história do NER é também, em grande medida, a história dos concursos de avaliação dedicados a essa tarefa. Os primeiros trabalhos sobre NER foram publicados na Conference on Message Understanding (MUC-6) (Grishman & Sundheim, 1995), que resultou numa tarefa partilhada destinada a identificar pessoas, lugares, organizações, expressões temporais e certos tipos de expressões numéricas em inglês. Posteriormente, surgiram inúmeros eventos análogos, como o CoNLL (neerlandês, espanhol, inglês e alemão) (Tjong Kim Sang, 2002; Tjong Kim Sang & De Meulder, 2003), o ACE (inglês, árabe e chinês) (Doddington et al., 2004) ou o HAREM (português) (Santos et al., 2006; Freitas et al., 2010), que também levaram ao desenvolvimento de corpora de avaliação para diferentes línguas e conjuntos de etiquetas.

Predominam duas abordagens:

- Baseadas em regras (geralmente expressões regulares, incluindo frequentemente informações gramaticais) e/ou dicionários de entidades (*gazetteers*). Por exemplo, Priberam (Amaral et al., 2008) e PALAVRAS (Bick, 2006) para português, e Linguakit (Gamallo & Garcia, 2017), com suporte para espanhol, galego, inglês e português.
- Baseadas em aprendizagem automática (*machine learning*), onde o conhecimento necessário para efetuar a tarefa é obtido a partir dos dados. Esta é a abordagem utilizada pelas ferramentas com suporte multilíngue mais conhecidas, como o OpenNLP[2], o CoreNLP (Manning et al., 2014) ou as seleccionadas como *baseline* neste trabalho.

As técnicas NER, por sua vez, têm sido aplicadas em múltiplos domínios e géneros, como a biomedicina (Oronoz et al., 2015; Giorgi & Bader, 2018), a microbiologia (Deléger et al., 2016), as redes sociais (Baldwin et al., 2015), a química (Eltyeb & Salim, 2014), a geologia (do Amaral

[1] https://spacy.io/

[2] https://opennlp.apache.org/

et al., 2017), o âmbito jurídico (Leitner et al., 2019; Pais et al., 2021) ou a análise literária, no contexto da leitura distante (Bamman et al., 2019; Frontini et al., 2020).

No domínio do turismo, onde se situa a nossa proposta, a identificação de entidades mencionadas também foi ganhando popularidade, embora outras técnicas como *topic modelling* ou a análise de sentimentos continuem a ser as mais utilizadas (Egger, 2022). A língua de trabalho é maioritariamente o inglês (Saputro et al., 2016; Vijay & Sridhar, 2016; Chantrapornchai & Tunsakul, 2019), embora também tenham sido desenvolvidos recursos e sistemas de NER para outras línguas, como o chinês (Guo et al., 2009; Xue et al., 2019), o mongol (Cheng et al., 2020), o português (Matos et al., 2021), o espanhol (García-Pablos et al., 2015) ou o árabe (Bouabdallaoui et al., 2022). Muitos dos sistemas produzidos utilizam corpora extraídos de avaliações de clientes em portais web ou redes sociais, mas não existem corpora anotados de entrevistas a visitantes, especialmente num subdomínio tão específico como o Caminho de Santiago.

2. Corpus

Para o treino e avaliação do sistema de reconhecimento de entidades mencionadas foi utilizado o subconjunto de textos em espanhol de CorGALE,[3] um corpus multilíngue desenvolvido a partir de entrevistas telefónicas a visitantes da cidade de Santiago de Compostela e anotado com as entidades *enamex* tradicionais introduzidas na CoNLL-2002 (Tjong Kim Sang, 2002): localização (LOC), pessoa (PER), organização (ORG) e miscelânea (MISC). O foco nos textos em espanhol deve-se ao facto de que, entre os sistemas selecionados como *baseline* (os *toolkits* Stanza, spaCy e Flair), apenas um tem modelos em português para NER (spaCy) e nenhum em galego.

A distribuição das entidades anotadas no corpus pode ver-se na Tabela 1, onde se constata um forte desequilíbrio a favor da categoria "localização," em proporções semelhantes para todas as línguas. Isto é de esperar devido à tipologia do corpus, que contém entrevistas em que um dos temas principais são os itinerários de viagem. Além disso, dos quatro tipos de entidades consideradas, a localização é a de maior interesse para o projeto em que se integra o presente trabalho, uma vez que um dos seus objetivos é a georreferenciação de entidades geográficas para a elaboração de mapas de densidade.

	subcorpus ES		**Total corpus**	
	freq.	%	freq.	%
LOC	18.387	84,98%	36.491	84,96%
MISC	1.606	7,42%	2.786	6,49%
PER	1.206	5,57%	2.639	6,14%
ORG	438	2,02%	1.037	2,41%
Total	21.637	100,00%	42.953	100,00%

Tabela 1: Distribuição de entidades no corpus.

Para além do forte desequilíbrio entre tipos de entidades, o domínio e o género (conversacional) do corpus condicionam métricas de diversidade lexical tais como a densidade de etiquetas, que mede a proporção de *tokens* que fazem parte de uma entidade mencionada em relação ao número total de *tokens*, ou a concentração de entidades, que representa o *ratio* entre entidades mencionadas (*NE tokens*, *named-entity tokens*) e entidades mencionadas únicas (*NE types*, *named-entity types*). Uma baixa densidade de etiquetas poderia influenciar negativamente a utilização do corpus como padrão de ouro para o treino de um modelo NER adaptado ao domínio.

Estas métricas de diversidade lexical podem ser contrastadas com as de outros corpora anotados para NER, de modo que as diferenças entre domínios e géneros distintos possam ser melhor apreciadas. Na primeira coluna da Tabela 2 apresentam-se estes valores de diversidade lexical em CorGALE. A fim de confrontar como estes valores são condicionados por diferenças de domínio e/ou género, adicionam-se também os relativos a outros corpora anotados com entidades mencionadas e um *tagset* semelhante. CoNLL e MET são corpora de género narrativo que incluem artigos jornalísticos em espanhol, enquanto os dois subcorpora ACE (Walker et al., 2006) são de género conversacional e em inglês. ACE BC contém transcrições de debates televisivos sobre notícias atuais e ACE CTS é um corpus de conversas telefónicas curtas entre dois participantes sobre um tópico (de uma lista total de 40) selecionado pelos investigadores. CoNLL tem, além disso, a particularidade de ter sido utilizado como corpus de treino para os modelos NER de Flair[4]

[3] Corpus GALabra de Entrevistas, que inclui entrevistas em galego, espanhol e português. O processo de compilação e anotação será publicado em breve: "Un corpus gold standard multilingüe para reconocimiento de entidades nombradas" https://galabra.socialdatalab.org/corgale

[4] https://huggingface.co/flair/ner-spanish-large

	CorGALE	ACE CTS	ACE BC	MET	CoNLL
língua género domínio	es conversacional turismo	en conversacional	en conversacional jornalístico	es narrativo jornalístico	es narrativo jornalístico
densidade de etiquetas	0,04	0,05	0,06	—	0,13
concentração de entidades	9,06	8,11	2,65	2,2	4,22

Tabela 2: Métricas de diversidade léxica do corpus em contraste com outros corpora comparáveis.

e Stanza,[5] utilizados aqui como *baseline*. No caso de CoNLL (Tjong Kim Sang, 2002) os cálculos foram feitos pelos autores. Os dados do corpus MET foram obtidos de Palmer & Day (1997), onde apenas se disponibiliza a concentração de entidades. Por fim, os dados de diversidade lexical do corpus ACE provêm de Augenstein et al. (2017, pp. 66–67).

Dos corpora comparados, os de domínio jornalístico apresentam valores muito inferiores de concentração de entidades, o que pode ser explicado pelo tipo de textos, porquanto costumam incluir um grande número de notícias que tratam de muitos tópicos diferentes, com o consequente aumento no número de entidades mencionadas que aparecem apenas uma vez. Por outro lado, os corpora do género conversacional tendem a apresentar uma menor densidade de etiquetas do que os de género narrativo, provavelmente condicionada por mecanismos de linguagem oral, tais como uma maior utilização de nexos, repetições ou palavras de preenchimento. Uma das implicações disto é que o CorGALE tem, em média, um número menor de entidades, e as que existem apresentam um alto número de repetições. Estas diferenças são relevantes, uma vez que os modelos NER generalistas são frequentemente treinados com corpora de género narrativo e domínio jornalístico, o que pode explicar um desempenho inferior desses modelos com outros tipos de textos.

3. Sistemas de Reconhecimento de Entidades Mencionadas

3.1. Baseline

Este trabalho utiliza os modelos NER das ferramentas generalistas Stanza, spaCy e Flair como *baseline*. Na Tabela 3 são apresentados os resultados da avaliação dessas ferramentas com o corpus CorGALE, permitindo assim determinar o nível de adaptação de cada uma delas ao domínio específico, que, como já vimos, condiciona (juntamente com o género) a distribuição das entidades. Para fins de contraste, também se incluíram os valores F1 declarados para esses modelos,[6] os quais foram produzidos utilizando um corpus de teste extraído do mesmo corpus usado para o treino. A comparação das duas avaliações mostra um baixo grau de adaptação ao domínio dos modelos, cujo desempenho se degrada quando são aplicados a domínios muito diferentes daqueles para os quais foram treinados. Isto reforça a necessidade de utilizar CorGALE para treinar um modelo NER específico, melhor adaptado ao domínio.

	Stanza	**spaCy**	**Flair**
modelo e versão	CoNLL02 (v.1.5.0)	es-core-news-lg (v.3.5.0)	ner-spanish-large (v.0.12.1)
CorGALI	0.686	0.744	0.752
	0.881	0.897	0.905

Tabela 3: *Baseline*: `f1-score` de Stanza, spaCy e Flair avaliados com CorGALE, em contraste com o `f1-score` declarado pelos modelos (última linha).

3.2. Experiências

De acordo com Liu et al. (2021), existem três paradigmas principais na evolução do proces-

[5] https://stanfordnlp.github.io/stanza/ner_models.html

[6] Os dados de desempenho do modelo NER de Stanza foram obtidos de Qi et al. (2020); os de spaCy provêm de https://github.com/explosion/spacy-models/releases/tag/es_core_news_lg-3.5.0 e os de Flair foram extraídos de https://huggingface.co/flair/ner-spanish-large.

samento de linguagem natural, que diferem na forma como os modelos aprendem a partir de um conjunto de dados: aprendizagem totalmente supervisionada (*fully supervised learning*), pré-treino e afinação (*pre-train and fine-tune*) e pré-treino, instrução e predição (*pre-train, prompt and predict*). Na aprendizagem totalmente supervisionada, o modelo é treinado através de um corpus de exemplos concebido para uma tarefa específica. Existem duas estratégias principais que podem ser utilizadas: a abordagem orientada às caraterísticas (*feature engineering*), na qual as caraterísticas são definidas manualmente com base no conhecimento do domínio, e a abordagem orientada à arquitetura (*architecture engineering*), na qual as caraterísticas relevantes são aprendidas automaticamente durante o treino do modelo utilizando arquiteturas de redes neuronais. Por sua vez, no paradigma de pré-treino e afinação, um modelo de linguagem com uma arquitetura fixa é treinado com grandes quantidades de dados não supervisionados, e depois adaptado a tarefas específicas por meio de afinação.

Neste trabalho, foram conduzidas experiências em sistemas NER utilizando um ou mais modelos para cada uma destas abordagens descritas. O corpus foi dividido aleatoriamente em duas partes, com 80% para treino e 20% para teste. Em todos os casos, utilizámos o mesmo corpus de treino e de teste, exceto no pré-treino e afinação, em que o corpus de treino foi dividido por sua vez em treino e validação.[7] Para a avaliação, não tomámos em consideração os valores produzidos pelos próprios modelos a fim de evitar possíveis problemas por diferenças no tratamento de sequências de etiquetas inadequadas (*improper label sequences* (Lignos & Kamyab, 2020)), nomeadamente aquelas que não correspondem a sequências permitidas pelo formato (por exemplo, uma etiqueta de tipo "I" após "O" em IOB2). Assim, utilizámos cada modelo para gerar predições a partir dos *tokens* do corpus de teste, e avaliámos essas predições em todos os casos usando `seqeval`,[8] uma biblioteca[9] de avaliação em Python que replica o comportamento de `conlleval`.[10]

Nas subsecções seguintes, descrevemos os modelos utilizados, classificados pela abordagem em que se inserem. Começamos apresentando os *Conditional Random Fields* (CRF) como parte da abordagem orientada às caraterísticas, seguido da abordagem orientada à arquitetura, na qual realizámos experiências com duas arquiteturas de redes neuronais: LSTM e CNN. Finalmente, no paradigma de pré-treino e *fine-tuning*, usamos duas bibliotecas para afinar seis modelos pré-treinados. Após a descrição dos modelos, expomos os resultados obtidos na sua avaliação, juntamente com uma primeira análise da capacidade de generalização. Por fim, fornecemos uma estimativa aproximada das emissões de carbono de cada experiência, bem como o tamanho final dos modelos produzidos.

3.3. *Feature engineering*: CRF

Os *Conditional Random Fields* (CRF, Lafferty et al. (2001)) são modelos probabilísticos de tipo discriminativo e com uma abordagem condicional, que definem as probabilidades de possíveis sucessões de etiquetas dada uma sequência observada. Na sua forma mais comum (*linear chain CRF*) são utilizados em aprendizagem automática para a anotação de dados sequenciais, como podem ser algumas tarefas de processamento de linguagem natural (etiquetação morfológica, reconhecimento de entidades mencionadas, *shallow parsing*) (Sha & Pereira, 2003), visão por computador (He et al., 2004) ou bioinformática (Settles, 2004; McDonald & Pereira, 2005).

Neste trabalho a implementação de CRF utilizada foi `sklearn-crfsuite`,[11] um *wrapper* Python da biblioteca CRFsuite[12] que oferece uma interface compatível com o `scikit-learn` (Pedregosa et al., 2011). O modelo foi treinado com 100 iterações, utilizando o algoritmo de optimização L-BFGS e os coeficientes de regularização L1 e L2 obtidos por `RandomizedSearch` durante a fase de optimização dos parâmetros. Foi realizada uma experiência sem limite de iterações, no qual a condição de paragem foi atingida na iteração 794. No entanto, o valor F1 do modelo final não apresentou melhoria em relação ao modelo de 100 iterações. No treino foram tidas en conta todas as transições,[13] de maneira que as impossíveis (por exemplo `O` → `I-LOC`) recebam

[7] Nesse caso, a proporção final foi de treino (60%), validação (20%) e teste (20%).

[8] O *script* de avaliação está disponível en https://github.com/sdocio/NER-experiments/blob/main/utils/eval.py

[9] seqeval, *A Python framework for sequence labeling evaluation*, https://github.com/chakki-works/seqeval.

[10] https://www.clips.uantwerpen.be/conll2000/chunking/conlleval.txt

[11] https://github.com/TeamHG-Memex/sklearn-crfsuite

[12] *CRFsuite: a fast implementation of Conditional Random Fields* http://www.chokkan.org/software/crfsuite/

[13] Opcão `all_possible_transitions=True`.

pesos negativos (em vez de zero), mesmo que não tenham sido observadas no corpus (Figura 1).

Uma das principais vantagens do CRF em relação a outros modelos sequenciais como HMM (*Hidden Markov Model*) é que não tem pressupostos de independência tão rigorosos e pode utilizar qualquer informação de contexto, juntamente com a possibilidade de usar um conjunto mais rico e flexível de caraterísticas (*features*). Para o nosso modelo[14] utilizámos uma série de *features* de tipo ortográfico[15] e contexto anterior e posterior. Foi também realizada uma experiência acrescentando *features* de tipo gramatical (etiquetas morfo-sintáticas), mas foi descartada uma vez que o desempenho do modelo era o mesmo.

3.4. *Architecture engineering*: redes neuronais

As redes neuronais (LeCun et al., 2015) são modelos computacionais inspirados no funcionamento dos cérebros biológicos, compostas de camadas de nodos interligados em que a saída de uma camada serve como entrada para a seguinte.

Para este trabalho, conduzimos experiências com duas arquiteturas de redes neuronais, *Bidirectional Long-Short Term Memory* combinado com CRF (BiLSTM-CRF) e *Convolutional Neural Networks* (CNN) com Bloom *embeddings*. Esta arquitetura é utilizada pelo spaCy num dos seus *pipelines*, o que se torna relevante por ser também um dos *baselines* usados. Em ambos os casos, durante o processo de treino utilizou-se o corpus supervisionado comum a todas as experimentações, juntamente com dados de corpora não supervisionados (nomeadamente, *embeddings* pré-treinados em corpora crus, sem etiquetas).

3.4.1. BiLSTM-CRF

As redes LSTM são um tipo de *Recurrent Neural Network* (RNN), uma família de redes neuronais adequada para trabalhar com sequências. No caso das redes neuronais bidirecionais, o modelo é composto por duas redes LSTM, uma que processa a sequência no sentido normal (para a frente) e outra que processa a sequência no sentido inverso (para trás). Essa abordagem permite capturar informação do contexto completo de cada elemento da sequência.

A arquitetura utilizada neste trabalho é a BiLSTM-CRF (Lample et al., 2016), em que a camada de entrada são *word embeddings* formados pela concatenação de duas classes de representações vetoriais. A primeira classe são os *character embeddings*, que capturam informações dos carateres que compõem cada palavra e são obtidos a partir do corpus de treino. A segunda classe são os *word-level embeddings*, que representam as palavras no seu contexto e são provenientes de dados externos não supervisionados (*embeddings* pré-treinados). Uma vez que no NER a sequência possível de etiquetas não é livre, em vez de modelar a etiquetagem assumindo probabilidades independentes, estas são modeladas em conjunto através de uma camada CRF.

Para o treino do nosso modelo, utilizamos uma adaptação da implementação `neural-ner`,[16] empregando *embeddings GloVe* (Pennington et al., 2014) com dimensão 300 para o espanhol,[17] que foram treinados com o corpus SBWC (Cardellino, 2019). Realizámos também testes com outros sistemas de representação vetorial, como o *Word2vec* e o *Fasttext*, que tiveram um desempenho inferior.

3.4.2. CNN com Bloom embeddings

As CNN são uma arquitetura de rede neuronal projetada especificamente para processar dados estruturados numa grelha. A implementação utilizada em esta arquitetura é o spaCy, que tem dois tipos de *pipeline* de processamento: o *pipeline* CPU com CNNs[18] e o *pipeline Transformer* (GPU).[19] O *pipeline* CPU integra como componente NER um modelo baseado em transições[20] inspirado no proposto por Lample et al. (2016, pp. 263–264). O spaCy adota a abordagem "Embed. Encode. Attend. Predict" (Honnibal, 2016), e emprega Redes Neuronais Convolucionais com conexões residuais (Residual CNNs) como codificador (*encoder*, na fase *Encode*). Além disso, utiliza *Bloom embeddings* (na fase *Embed*), que reduzem o tamanho da tabela de vetores por meio de quatro funções de *hash* e combinam caraterísticas ortográficas, como NORM (palavra normalizada), PREFIX (prefixo), SUFFIX (sufixo) e SHAPE (forma da palavra) (Miranda et al., 2022; Honnibal et al., 2022).

[14] https://github.com/sdocio/NER-experiments/tree/main/0-crf

[15] Por exemplo, se uma palavra começa com uma letra maiúscula ou contém números.

[16] https://github.com/yuhaozhang/neural-ner

[17] https://github.com/dccuchile/spanish-word-embeddings

[18] https://spacy.io/models#design-cnn

[19] https://spacy.io/models#design-trf

[20] https://spacy.io/api/architectures#parser

From \ To	O	B-LOC	I-LOC	B-MISC	I-MISC	B-ORG	I-ORG	B-PER	I-PER
O	5.847	3.113	-5.29	2.996	-3.909	2.094	-3.921	1.745	-3.268
B-LOC	-0.017	-2.102	6.687	-1.464	-2.62	-0.862	-1.551	-1.928	-1.821
I-LOC	-0.932	-3.834	5.993	-1.248	-1.395	-0.8	-1.775	-1.665	-1.296
B-MISC	-0.687	-2.657	-2.364	-1.789	7.558	-0.7	-1.69	-1.443	-1.235
I-MISC	-1.379	-3.589	-3.067	-1.479	6.897	-1.402	-1.895	-3.115	-1.748
B-ORG	-1.724	-1.8	-1.814	-1.132	-0.71	-1.643	5.786	-1.458	-0.663
I-ORG	-1.338	-3.493	-1.92	-0.849	-1.032	-0.855	6.088	-1.231	-0.592
B-PER	-0.451	-1.917	-1.772	-0.809	-1.114	-0.522	-1.1	-2.557	6.117
I-PER	-0.144	-1.113	-1.148	-0.638	-0.452	-0.397	-0.483	-1.606	5.564

Figura 1: Pesos das transições.

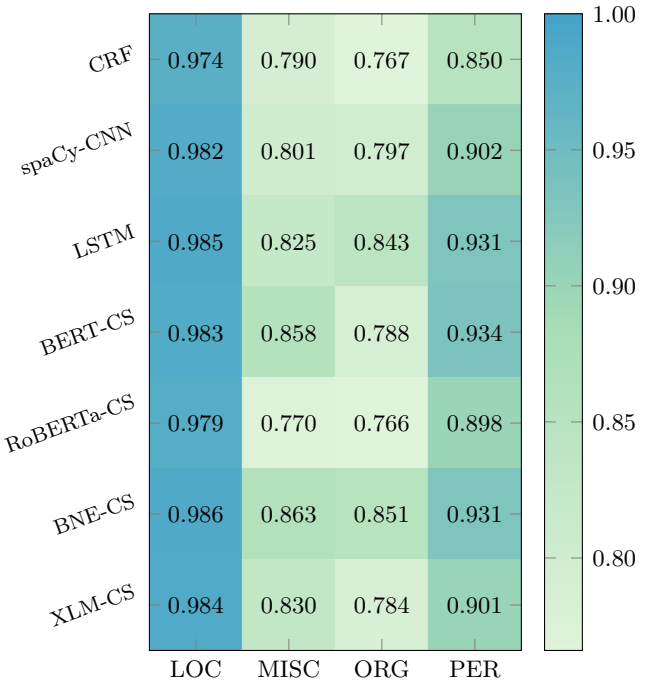

Figura 2: Valores `F1` de todos os experimentos divididos por etiqueta.

3.5. *Objective engineering*: afinação de *Transformers*

Um *Transformer* é uma arquitetura de rede neuronal baseada em mecanismos de atenção (Vaswani et al., 2017). Nessa arquitetura, os mecanismos de atenção substituem o processamento sequencial das Redes Neuronais Recorrentes (RNN, *Recurrent Neural Networks*), difícil de paralelizar, por um processamento distribuído altamente eficiente e paralelizável, concebido para dar conta de relações de palavras a longa distância. A principal vantagem dos *Transformers* é o facto de permitirem gerar modelos de língua genéricos que podem ser ajustados a qualquer tarefa de jeito rápido e eficiente através da afinação (*fine-tuning*) dos parâmetros (Devlin et al., 2019). A afinação é uma técnica de aprendizagem por transferência que parte de um grande modelo de língua (LLM, *Large Language Model*) treinado mediante aprendizagem auto-supervisionada (modelo pré-treinado). Esses grandes modelos pré-treinados utilizam a predição de palavras em contexto previamente mascaradas para aprender, sem a necessidade de anotação manual. Depois, o modelo pré-treinado é adaptado para uma tarefa específica através da transferência de seus parâmetros, ajustando-os com um pequeno *dataset* anotado. Esse processo supervisionado gera um novo modelo afinado, que herda o mesmo número de parâmetros do modelo pré-treinado genérico.

Para este trabalho, os modelos afinados para NER aplicado ao nosso domínio foram:

ner-cds-bert (BERT-CS): Foi afinado utilizando como modelo pré-treinado BETO (Cañete et al., 2020), um modelo baseado em BERT (Devlin et al., 2019) e produzido utilizando um corpus de três mil milhões de palavras provenientes de diversas fontes (Cañete, 2019).

ner-cds-spanberta (RoBERTa-CS): Foi ajustado utilizando como modelo SpanBERTa,[21] baseado em RoBERTa (Liu et al., 2019) e que foi treinado com a parte em espanhol do corpus OSCAR (Ortiz Suárez et al., 2019).

ner-cds-bne (BNE-CS): Foi afinado utilizando como modelo pré-treinado RoBERTa-Base-BNE (Gutiérrez Fandiño et al., 2022), um modelo de língua baseado na arquitetura de RoBERTa e pré-treinado usando um corpus de espanhol com 135 mil millões de palavras, que foi compilado a partir do arquivo web construído pela *Biblioteca Nacional de España* de 2009 a 2019. Para a realização das experiências, foram utilizadas duas

[21] https://github.com/chriskhanhtran/spanish-bert

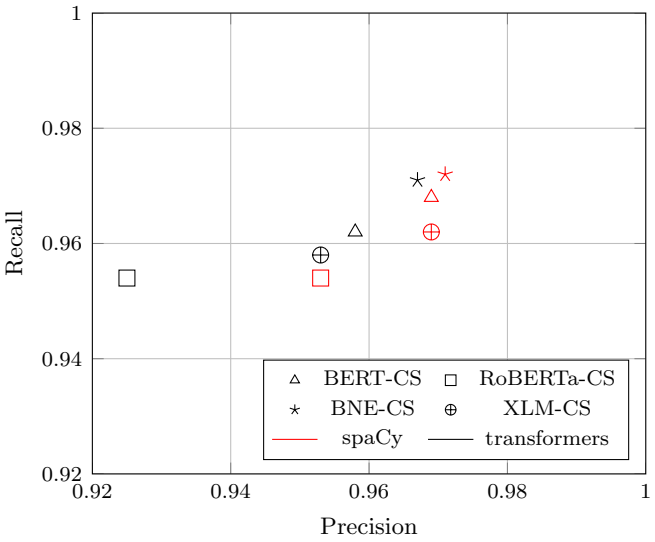

Figura 3: Comparação dos resultados da afinação com as bibliotecas *spaCy* e *transformers*.

variantes: a *base* e a *large*. No entanto, a variante *large* foi rejeitada por ter apresentado piores resultados.

ner-cds-xlm (XLM-CS): Baseado no modelo XLM-RoBERTa-Large (Conneau et al., 2019), uma versão multilíngue de RoBERTa que foi pré-treinado com dados do corpus CommonCrawl em 100 línguas. A variante *base* também foi testada, mas foi descartada por apresentar resultados inferiores.

Os modelos selecionados para esta afinação são todos os LLMs de auto-codificação (*autoencoding*) de livre acesso, disponíveis para espanhol e citados na literatura. Para ajustar esses modelos, foram utilizadas duas bibliotecas: *transformers* (Wolf et al., 2020) de HuggingFace (versão 4.28.1), e spaCy de Explosion (versão 3.5.2).

A Figura 3 apresenta uma comparação dos resultados do processo de afinação utilizando ambas as bibliotecas. De forma geral, os modelos afinados com a biblioteca spaCy têm melhor desempenho do que os afinados com a biblioteca *transformers*, pelo que, para o confronto final na Tabela 4, apenas consideraremos os primeiros.

4. Resultados

Na Tabela 4 estão apresentados os valores de *precision*, *recall* e `F1` obtidos na avaliação para cada uma das abordagens testadas, juntamente com os resultados dos modelos *baseline*. Podemos observar que o modelo com melhor desempenho (BNE-CS) corresponde ao paradigma de *fine-tuning* de *Transformers*, e supera em mais de 20 pontos os sistemas pré-treinados utilizados como ponto de partida. Embora BNE-CS apresente o melhor desempenho, o modelo XLM-CS está muito próximo em relação a todos os valores medidos. Este modelo tem a particularidade de ser o resultado da afinação de um modelo multilíngue, o que pode torná-lo adequado para as três línguas do corpus CorGALE.

Por outra parte, o mapa de calor da Figura 2 mostra o valor `F1` segmentado por classe, revelando um padrão semelhante em todos os modelos treinados, com resultados superiores para as etiquetas LOC e PER, enquanto apresentam um desempenho inferior nas classes MISC e ORG. É importante destacar que a etiqueta ORG é a menos representada no corpus (Tabela 2).

Finalmente, a matriz de confusão representada na Figura 4 ilustra a relação entre as previsões (eixo X) e os valores reais (eixo Y) do modelo com melhor desempenho. Foi produzida comparando o corpus de teste com as previsões do modelo.[22] A maior intensidade de cor na diagonal indica uma taxa de acerto significativa em todas as classes. Além disso, a figura evidencia que as taxas de erro mais altas estão relacionadas com problemas na delimitação de entidades nas classes LOC e ORG (confusão entre B-LOC e I-LOC e entre B-ORG e I-ORG).

4.1. Análise da capacidade de generalização dos modelos

Para o treino e avaliação dos modelos, seguimos a prática comum de dividir aleatoriamente o conjunto de dados em corpora de treino e de teste, conforme descrito na secção 3.[23] No entanto, esse procedimento pode estar a sobrestimar o desempenho real dos modelos, uma vez que um valor elevado numa avaliação não implica necessariamente uma verdadeira generalização (Kim & Kang, 2022), ou seja, a capacidade de um modelo para fazer previsões precisas sobre dados não observados durante o treino. No nosso caso, os resultados podem ser atribuídos, em parte, à alta concentração de entidades no corpus (ver Tabela 2) e ao facto de os modelos estarem, em boa medida, a aprender as entidades vistas no corpus de treino (memorização), resultando num peso menor da interpretação do contexto linguístico na predição da etiqueta (Agarwal et al., 2021).

[22] Script utilizado: https://github.com/sdocio/NER-experiments/blob/main/utils/matrix.py

[23] Foi realizada uma experiência para testar o modelo CRF (secção 3.3) utilizando KFold Cross Validation (K=10). Os resultados não divergem dos obtidos com a divisão clássica 80%-20%, com um valor F1 de 0,952 e um desvio padrão de 0,004. Testes semelhantes com os restantes modelos foram descartados devido ao seu elevado custo de computação.

Paradigma		Abordagem	Modelo	Precision	recall	f1-score
Modelos treinados	Aprendizagem totalmente supervisionada	eng. orientada a caraterísticas (*feature engineering*)	CRF	0.955	0.947	0.951
		eng. orientada à arquitetura (*architecture engineering*)	spaCy-CNN	0.963	0.959	0.961
			BiLSTM-CRF	0.969	0.967	0.968
	Pré-treino e afinação (*fine-tuning*)	eng. orientada a objetivos (*objective engineering*)	RoBERTa-CS	0.953	0.954	0.953
			BERT-CS	0.969	0.968	0.968
			BNE-CS	**0.971**	**0.972**	**0.971**
			XLM-CS	0.969	0.962	0.965
Baseline			Stanza	0.683	0.689	0.686
			spaCy	0.683	0.816	0.744
			Flair	0.748	0.756	0.752

Tabela 4: Comparação final com todos os modelos produzidos.

Para determinar de maneira mais eficaz a capacidade de generalização dos nossos modelos, efetuámos uma primeira análise preliminar, avaliando o desempenho dos modelos com entidades do corpus de teste que não estão presentes no corpus de treino.

Para realizar esse procedimento, conduzimos um teste no qual dividimos os segmentos contendo entidades mencionadas do corpus de teste em duas partes distintas:

- *Corpus de entidades não observadas*: consiste exclusivamente nos segmentos que contêm entidades que não foram observadas no corpus de treino.
- *Corpus de entidades observadas*: compreende aqueles segmentos que possuem uma ou mais entidades observadas no corpus de treino. Nesta parte, também podem estar presentes entidades não observadas que compartilham o segmento com as entidades observadas.

Os segmentos sem entidades, ou seja, as orações que não possuem nenhuma anotação, foram divididos equitativamente entre as duas partes.

A Tabela 5 mostra os resultados dos diferentes modelos com o corpus de entidades não observadas, apresentando o valor do `F1` resultante e a percentagem de variação em relação ao `F1` obtido com o corpus de teste completo. Todos os modelos evidenciam uma diminuição acentuada nos valores de `F1`, com uma média de 30.78%. O modelo CRF registou a maior descida com 52.37%, o que é de esperar, tendo em conta que a informação contextual de que dispõe é muito limitada. Por outro lado, o modelo com o melhor desempenho no corpus de teste original (BNE-CS) é também o melhor na prova com as entidades não observadas, apresentando uma redução de 18,27% na versão afinada com a biblioteca *transformers* e de 22,86% na versão afinada com a biblioteca *spaCy*. Os resultados apresentados neste

	modelo	f1-score	% variação
	CRF	**0.453**	**-52.37%**
	spaCy-CNN	0.656	-31.74%
	LSTM-CRF	0.684	-29.34%
spaCy	BERT-CS	0.714	-26.24%
	RoBERTa-CS	0.662	-30.54%
	BNE-CS	0.749	-22.86%
	BNE-CS lg	0.523	-45.18%
	XLM-CS	0.646	-32.85%
	XLM-CS lg	0.660	-31.61%
transformers	BERT-CS	0.680	-29.17%
	RoBERTa-CS	0.663	-29.39%
	BNE-CS	**0.792**	**-18.27%**
	BNE-CS lg	0.700	-26.93%
	XLM-CS	0.699	-26.81%
	XLM-CS lg	0.684	-28.38%

Tabela 5: Valores `f1-score` para o corpus de entidades não observadas no treino. O número de entidades neste corpus é de 193, muito inferior às 4265 do corpus de teste completo.

teste são próximos dos obtidos noutros testes semelhantes, como o conduzido por Kádár et al. (2023).

É interessante destacar que, enquanto na avaliação com o corpus de teste completo os modelos afinados com o *spaCy* tiveram melhor desempenho do que os afinados com a biblioteca *transformers* (ver Figura 3), na avaliação com o corpus de entidades não observadas durante o treino o resultado é o oposto em todos os casos, exceto para o modelo BERT-CS.

4.2. Discussão

Os resultados obtidos mostram uma acurácia elevada, que, no caso do melhor modelo, é de 0.998. Como se pode ver, superam não apenas os sistemas de referência avaliados com o corpus CorGALE (Tabela 3), mas também o desempenho de outros modelos NER para espanhol treinados com corpora jornalísticos e etiquetas comparáveis (Tabela 6).

modelo	f1-score
bert-spanish-cased-finetuned-ner	90.17
bertin-base-ner-conll2002-es	87.06
bne-spacy-corgale-ner-es (BNE-CS)	**97.13**
ner-spanish-large	90.54
roberta-large-bne-capitel-ner	90.51
xlm-roberta-large-ner-spanish	89.17

Tabela 6: Resultados de modelos NER para espanhol

Porém, os testes realizados demonstram que, embora as avaliações com uma divisão clássica de corpus aleatório possam indicar, em termos gerais, quais modelos apresentam melhor desempenho no domínio de trabalho, não são suficientes para obter uma compreensão clara de sua verdadeira generalização. Portanto, é necessário substituir a divisão aleatória do corpus por divisões baseadas em heurísticas com diferentes enviesamentos ou adotar uma abordagem de divisão de dados usando um desenho adversarial (Søgaard et al., 2021).

Para além disso, o sistema tem claras limitações, como o forte desequilíbrio entre as classes ou a elevada concentração de entidades mencionadas no corpus de treino, o que pode estar a condicionar os resultados obtidos.

4.3. Emissão de CO_2 e tamanho dos modelos

Outros fatores importantes a serem considerados na avaliação dos resultados de um modelo são o tamanho e a pegada de carbono decorrente do seu treino, expressa em kg CO_2eq. Para calcular essa pegada de carbono, utilizamos a fórmula proposta por Lacoste et al. (2019):

$$\left(\frac{W \times t}{1000} \times E\right) \times PUE \quad (1)$$

onde W é o consumo de energia do hardware, t é o tempo em horas, E representa as emissões médias de carbono da rede de energia utilizada e PUE (*Power Usage Effectiveness*) contabiliza a energia adicional necessária para sustentar a infraestrutura de computação, principalmente a refrigeração.

As experiências foram realizadas numa GPU Nvidia A100 PCIe-40GB, com um consumo de energia declarado de 250W e conetada a um centro de processamento de dados local. O valor utilizado para E é de 0,2120 kg CO_2eq/kWh, que representa a emissão média de dióxido de carbono por quilowatt-hora em Espanha para o ano 2022.[24] O coeficiente PUE aplicado é 1.58, a média global dos centros de dados em 2018, conforme mencionado por Strubell et al. (2020). No entanto, o modelo CRF foi treinado com uma CPU Intel® Core™ i7-10510U e um consumo aproximado de energia durante o treino de 19W. Neste caso, o coeficiente PUE aplicado é 1, uma vez que o consumo de energia já inclui tanto a computação quanto a infraestrutura.

A Tabela 7 apresenta as emissões de carbono produzidas durante o treino dos modelos (em kg CO_2eq) e seus respetivos tamanhos em megabytes, exibindo uma grande variabilidade em ambos os valores. O total das emissões é de aproximadamente um quilograma e meio de CO_2eq, o que não inclui o custo de produção dos modelos pré-treinados, como por exemplo, os LLMs no caso de *fine-tuning* ou os modelos de *embeddings*. As emissões médias são de 0,085 kg CO_2eq, sendo o CRF o modelo mais eficiente em termos energéticos, com uma redução de cerca de 99.93% em relação à média, e também o de menor tamanho. O LSTM-CRF é o menos eficiente em termos energéticos (por volta de 234% acima da média). No caso do modelo com melhor desempenho, o BNE-CS afinado com a biblioteca spaCy apresenta umas emissões estimadas de 0.045, cerca de 47.57% abaixo da média.

[24] https://app.electricitymaps.com/zone/ES

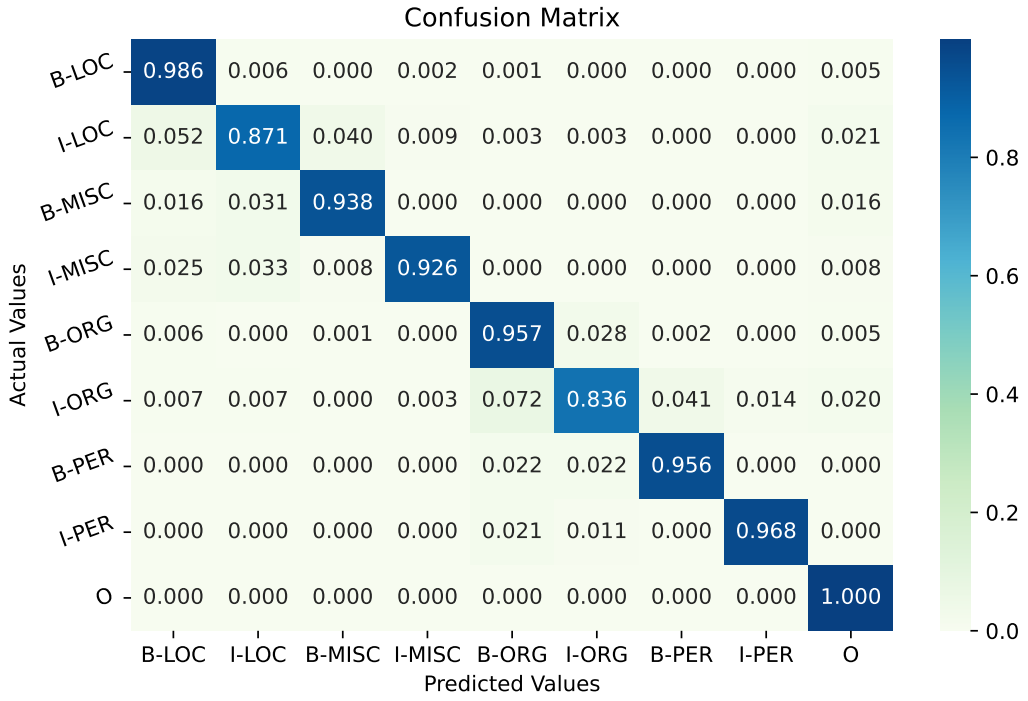

Figura 4: Matriz de confusão do melhor modelo: *BNE-CS* afinado com spaCy.

	modelo	horas	kg.CO_2eq	tamanho
	CRF	**0.01**	**0.0001**	**0.848 MB**
	spaCy-CNN	0.23	0.0196	622 MB
	LSTM GloVe	3.39	0.2839	24 MB
	LSTM Fasttext	3.38	0.2830	23 MB
	LSTM Word2vec	**3.40**	**0.2851**	**23 MB**
spaCy	BERT-CS	0.56	0.0470	421 MB
spaCy	RoBERTa-CS	0.40	0.0331	480 MB
spaCy	BNE-CS	0.53	0.0445	480 MB
spaCy	BNE-CS lg	1.45	0.1215	1360 MB
spaCy	XLM-CS	1.21	0.1017	1083 MB
spaCy	XLM-CS lg	1.42	0.1190	2158 MB
transformers	BERT-CS	0.14	0.0120	418 MB
transformers	RoBERTa-CS	0.14	0.0117	477 MB
transformers	BNE-CS	0.15	0.0122	477 MB
transformers	BNE-CS lg	0.30	0.0254	1356 MB
transformers	XLM-CS	0.16	0.0135	1075 MB
transformers	XLM-CS lg	0.35	0.0296	2149 MB
	Total	**17.24**	**1.4428**	

Tabela 7: Pegada de carbono do treino dos modelos produzidos.

Em relação ao tamanho, é de salientar que existe uma grande diferença entre os modelos maiores e mais pequenos, em contraste com os valores muito próximos de `F1` obtidos na avaliação. Este facto pode influenciar a seleção entre um ou outro modelo em ambientes de produção, tornando preferíveis os modelos mais pequenos com menos emissões, mesmo que tenham desempenhos ligeiramente inferiores.

5. Conclusões e trabalho futuro

Neste artigo, apresentamos uma avaliação de diversas abordagens para desenvolver um modelo de reconhecimento de entidades mencionadas aplicado ao domínio do turismo e da análise cultural em espanhol. Também descrevemos brevemente o recurso utilizado para avaliar e treinar os modelos, um corpus de entrevistas telefónicas anotado com o objetivo de servir como padrão de ouro para NER. Foram introduzidas algumas considerações sobre a diversidade das entidades mencionadas e sobre a forma como esta é afetada tanto pelo domínio do corpus como pelo seu género.

A seguir, utilizámos o corpus para avaliar várias ferramentas NER pré-treinadas e mostrámos que os resultados obtidos não são satisfatórios. Para resolver este problema, foram treinados vários modelos adaptados a este domínio, utilizando diferentes abordagens que vão desde os algoritmos clássicos de aprendizagem automática até a afinação de *Transformers*. Os resultados das experiências realizadas melhoram amplamente o ponto de partida (*baseline*)

das ferramentas pré-treinadas avaliadas. No entanto, as avaliações realizadas não indicam necessariamente um bom desempenho na generalização do que foi aprendido. Alguns testes preliminares mostram que, quando o modelo é confrontado com entidades não observadas durante o treino, o seu rendimento diminui significativamente. O próximo passo será dividir o corpus em conjuntos de treino e teste utilizando um esquema de teste adversarial (em vez do método aleatório convencional), bem como efetuar avaliações adicionais usando as ampliações planeadas para o corpus.

Outra questão a ser explorada no futuro será a relacionada com o terceiro paradigma, "pré-treino, instrução e predição"(*pre-train, prompt and predict*), em que não há uma adaptação à tarefa de um modelo de língua pré-treinado. Em vez disso, são utilizadas instruções adequadas (*prompts*) para fazer com que um modelo linguístico pré-treinado produza as predições desejadas sem a necessidade de um treino específico (tornando-o completamente não supervisionado). No nosso caso, foram realizadas provas preliminares com o modelo GPT-3 (`text-davinci-003`), utilizando um pequeno subconjunto do corpus de teste (aproximadamente 10% do total, selecionado aleatoriamente) e um *prompt* incluindo dois exemplos de etiquetação. Os resultados obtidos são inferiores em relação ao *baseline*, além de apresentar outros problemas, como modificações introduzidas no texto resultante ou a presença de etiquetas que não faziam parte do conjunto de etiquetas. No entanto, é necessário realizar testes mais abrangentes com um conjunto maior de *prompts*, utilizando outros modelos de língua (como o LLaMA[25] ou o BLOOM[26]) e com um corpus de teste completo para que os resultados sejam comparáveis às restantes experiências.

Por último, como o corpus utilizado, CorGALE, é multilíngue, uma outra linha de trabalho consistirá em replicar as experiências com os subcorpora galego e português, e prestar atenção especial aos resultados oferecidos por modelos multilíngues, como o XLM-RoBERTa.[27]

Agradecimentos

Este trabalho faz parte do projeto *Narrativas, usos e consumo dos visitantes como aliados ou ameaças ao bem-estar da comunidade local: o caso de Santiago de Compostela*, com referência FFI2017-88196-R, parcialmente subsidiado pela Agencia Estatal de Investigación (AEI) - Fundos Feder (de janeiro de 2018 a junho de 2022).

As experiências foram realizadas nos Clusters de Computação de Alto Desempenho (HPC) do Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) da Universidade de Santiago de Compostela e do Centro de Supercomputación de Galicia (CESGA).

Os autores agradecem ainda ao editor e aos revisores da Linguamática a revisão e comentários, que ajudaram a melhorar o artigo.

Tanto o código como os modelos treinados descritos neste trabalho estão disponíveis:

- https://github.com/sdocio/NER-experiments
- https://huggingface.co/sdocio.

Referências

Agarwal, Oshin, Yinfei Yang, Byron C. Wallace & Ani Nenkova. 2021. Interpretability analysis for named entity recognition to understand system predictions and how they can improve. *Computational Linguistics* 47(1). 117–140. 10.1162/coli_a_00397.

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter & Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art nlp. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 54–59. 10.18653/v1/N19-4010.

Amaral, Carlos, Helena Figueira, Afonso Mendes, Pedro Mendes, Cláudia Pinto & Tiago Veiga. 2008. Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 171–179. Linguateca.

Augenstein, Isabelle, Leon Derczynski & Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language* 44. 61–83. 10.1016/j.csl.2017.01.012.

Baldwin, Timothy, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter & Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition.

[25] https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

[26] https://huggingface.co/bigscience/bloom

[27] https://huggingface.co/docs/transformers/model_doc/xlm-roberta

Em *Workshop on Noisy User-generated Text*, 10.18653/v1/W15-4319.

Bamman, David, Sejal Popat & Sheng Shen. 2019. An annotated dataset of literary entities. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2138–2144. 10.18653/v1/N19-1220.

Barachi, May El, Sujith Samuel Mathew & Manar AlKhatib. 2022. Combining named entity recognition and emotion analysis of tweets for early warning of violent actions. Em *7th International Conference on Smart and Sustainable Technologies (SpliTech)*, 1–6. 10.23919/SpliTech55088.2022.9854231.

Bick, Eckhard. 2006. Functional aspects in Portuguese NER. Em *Computational Processing of the Portuguese Language (PROPOR)*, 80–89.

Bouabdallaoui, Ibrahim, Fatima Guerouate, Samya Bouhaddour, Chaimae Saadi & Mohamed Sbihi. 2022. Named entity recognition applied on Moroccan tourism corpus. Em *12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare*, vol. 198, 373–378. 10.1016/j.procs.2021.12.256.

Cañete, José. 2019. Compilation of large Spanish unannotated corpora. Version 2. Zenodo. 10.5281/zenodo.3247731.

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang & Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. Em *Practical Machine Learning for Developing Countries at ICLR*, s.p.

Cardellino, Cristian. 2019. Spanish billion words corpus and embeddings. https://crscardellino.github.io/SBWCE/.

Chantrapornchai, Chantana & Aphisit Tunsakul. 2019. Information extraction based on named entity for tourism corpus. Em *16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 187–192. 10.1109/JCSSE.2019.8864166.

Cheng, Xiao, Weihua Wang, Feilong Bao & Guanglai Gao. 2020. MTNER: A corpus for Mongolian tourism named entity recognition. Em Junhui Li & Andy Way (eds.), *Machine Translation*, 11–23. 10.1007/978-981-33-6162-1_2.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR* abs/1911.02116. http://arxiv.org/abs/1911.02116.

Deléger, Louise, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières & Claire Nédellec. 2016. Overview of the bacteria biotope task at BioNLP shared task. Em *4th BioNLP Shared Task Workshop*, 12–22. 10.18653/v1/W16-3002.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. 10.18653/v1/N19-1423.

do Amaral, Daniela O. F., Sandra Collovini, A. Figueira, Renata Vieira & Marco Gonzalez. 2017. Processo de construção de um corpus anotado com entidades geológicas visando REN. Em *11th Brazilian Symposium in Information and Human Language Technology*, 63–72.

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel & Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 837–840.

Egger, Roman (ed.). 2022. *Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications* Tourism on the Verge. Cham: Springer International Publishing. 10.1007/978-3-030-88389-8.

Eltyeb, Safaa & Naomie Salim. 2014. Chemical named entities recognition: a review on approaches and applications. *Journal of Cheminformatics* 6. 17. 10.1186/1758-2946-6-17.

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.

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos &

Ranka Stanković. 2020. Named entity recognition for distant reading in ELTeC. Em *CLARIN Annual Conference*, 37–41.

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.

García-Pablos, Aitor, Montse Cuadros & Maria Teresa Linaza. 2015. OpeNER: Open tools to perform natural language processing on accommodation reviews. Em *Information and Communication Technologies in Tourism*, 125–137. 10.1007/978-3-319-14343-9_10.

Giorgi, John M. & Gary D. Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics* 34(23). 4087–4094. 10.1093/bioinformatics/bty449.

Grishman, Ralph & Beth Sundheim. 1995. Design of the MUC-6 evaluation. Em *6th Conference on Message Understanding*, 1–11. 10.3115/1072399.1072401.

Guo, Jianyi, Zhengshan Xue, Zhengtao Yu, Zhikun Zhang, Yihao Zhang & Xianming Yao. 2009. Named entity recognition for the tourism domain based on cascaded conditional random fields. *Journal of Chinese Information Processing* 23(5). 47–52.

Gutiérrez Fandiño, Asier, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio-Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre & Marta Villegas. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural* 68. 39–60. 10.26342/2022-68-3.

He, Xuming, Richard S. Zemel & Miguel A. Carreira-Perpiñán. 2004. Multiscale conditional random fields for image labeling. Em *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, II–II. 10.1109/CVPR.2004.1315232.

Honnibal, Matthew. 2016. Embed, encode, attend, predict: The new deep learning formula for state-of-the-art NLP models. Explosion. https://explosion.ai/blog/deep-learning-formula-nlp.

Honnibal, Matthew, Adriane Boyd & Vincent D. Warmerdam. 2022. Compact word vectors with bloom embeddings. Explosion. https://explosion.ai/blog/bloom-embeddings.

Kanev, Anton I., Grigory A. Savchenko, Ilya A. Grishin, Denis A. Vasiliev & Emilia M. Duma. 2022. Sentiment analysis of multilingual texts using machine learning methods. Em *Conference of Russian Young Researchers in Electrical and Electronic Engineering*, 326–331. 10.1109/ElConRus54750.2022.9755568.

Kim, Hyunjae & Jaewoo Kang. 2022. How do your biomedical named entity recognition models generalize to novel entities? *IEEE Access* 10. 31513–31523. 10.1109/ACCESS.2022.3157854.

Kádár, Ákos, Lester James Miranda, Victoria Slocum & Sofie Van Landeghem. 2023. The tale of bloom embeddings and unseen entities. Explosion. https://explosion.ai/blog/technical-report.

Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt & Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. ArXiv [cs.CY]. 10.48550/ARXIV.1910.09700.

Lafferty, John, Andrew McCallum & Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Em *18th International Conference on Machine Learning*, 282–289.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami & Chris Dyer. 2016. Neural architectures for named entity recognition. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270. 10.18653/v1/N16-1030.

LeCun, Yann, Yoshua Bengio & Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553). 436–444. 10.1038/nature14539.

Lee, Jangwon, Jungi Lee, Minho Lee & Gil-Jin Jang. 2022. Named entity correction in neural machine translation using the attention alignment map. *Applied Sciences* 11(15). 10.3390/app11157026.

Leitner, Elena, Georg Rehm & Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. Em *15th International Conference, SEMANTiCS*, 272–287. 10.1007/978-3-030-33220-4.

Lignos, Constantine & Marjan Kamyab. 2020. If you build your own NER scorer, non-replicable results will come. Em *1st Workshop on Insights from Negative Results in NLP*, 94–99. 10.18653/v1/2020.insights-1.15.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi & Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55. 1–35. 10.1145/3560815.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv [cs.CL]. 10.48550/arXiv.1907.11692.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. Em *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60. 10.3115/v1/P14-5010.

Matos, Emanuel, Mário Rodrigues, Pedro Miguel & António Teixeira. 2021. Towards automatic creation of annotations to foster development of named entity recognizers. Em *10th Symposium on Languages, Applications and Technologies (SLATE)*, vol. 94, 11:1–11:14. 10.4230/OASIcs.SLATE.2021.11.

McDonald, Ryan & Fernando Pereira. 2005. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics* 6(Suppl 1). S6. 10.1186/1471-2105-6-S1-S6.

Miranda, Lester James, Ákos Kádár, Adriane Boyd, Sofie Van Landeghem, Anders Søgaard & Matthew Honnibal. 2022. Multi hash embeddings in spaCy. ArXiv [cs.CL]. 10.48550/arXiv.2212.09255.

Oronoz, Maite, Koldo Gojenola, Alicia Pérez, Arantza Díaz de Ilarraza & Arantza Casillas. 2015. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics* 56. 318–332. 10.1016/j.jbi.2015.06.016.

Ortiz Suárez, Pedro Javier, Benoît Sagot & Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Em *Workshop on Challenges in the Management of Large Corpora*, 9–16. 10.14618/ids-pub-9021.

Pais, Vasile, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi & Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. Em *Natural Legal Language Processing Workshop*, 9–18. 10.18653/v1/2021.nllp-1.2.

Palmer, David D. & David S. Day. 1997. A statistical profile of the named entity task. Em *5th Conference Applied Natural Language Processing*, 190–193. 10.3115/974557.974585.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

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

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. Em *58th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, 101–108. 10.18653/v1/2020.acl-demos.14.

Santos, Diana, Nuno Seco, Nuno Cardoso & Rui Vilela. 2006. HAREM: An advanced NER evaluation contest for Portuguese. Em *5th International Conference on Language Resources and Evaluation (LREC)*, 1986–1991.

Saputro, Khurniawan Eko, Sri Suning Kusumawardani & Silmi Fauziati. 2016. Development of semi-supervised named entity recognition to discover new tourism places. Em *2nd International Conference on Science and Technology-Computer (ICST)*, 124–128. 10.1109/ICSTC.2016.7877360.

Settles, Burr. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. Em *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 107–110.

Sha, Fei & Fernando Pereira. 2003. Shallow parsing with conditional random fields. Em *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 213–220.

Søgaard, Anders, Sebastian Ebert, Jasmijn Bastings & Katja Filippova. 2021. We need to talk about random splits. Em *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1823–1832. 10.18653/v1/2021.eacl-main.156.

Strubell, Emma, Ananya Ganesh & Andrew McCallum. 2020. Energy and Policy Considerations for Modern Deep Learning Research. Em *AAAI Conference on Artificial Intelligence*, vol. 34 9, 13693–13696. 10.1609/aaai.v34i09.7123.

Tjong Kim Sang, Erik F. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. Em *6th Conference on Natural Language Learning (CoNLL)*, s.p.

Tjong Kim Sang, Erik F. & Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Em *7th Conference on Natural Language Learning (CoNLL)*, 142–147.

Torres Feijó, Elias J. 2019. *Bem-estar comunitário e visitantes através do Caminho de Santiago. Grandes narrativas, ideias e práticas culturais na cidade*. Andavira.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. Em *31st Annual Conference on Neural Information Processing Systems*, vol. 1, 5999–6008.

Vijay, J. & Rajeswari Sridhar. 2016. A machine learning approach to named entity recognition for the travel and tourism domain. *Asian Journal of Information Technology* 15(21). 4309–4317. 10.3923/ajit.2016.4309.4317.

Vu, Van-Hai, Quang-Phuoc Nguyen, Kiem-Hieu Nguyen, Joon-Choul Shin & Cheol-Young Ock. 2020. Korean-Vietnamese Neural Machine Translation with Named Entity Recognition and Part-of-Speech Tags. *IEICE Transactions on Information and Systems* E103.D(4). 866–873. 10.1587/transinf.2019EDP7154.

Walker, Christopher, Stephanie Strassel, Julie Medero & Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. Linguistic Data Consortium. 10.35111/mwxc-vh88.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. 10.18653/v1/2020.emnlp-demos.6.

Xue, Leyi, Han Cao, Fan Ye & Yuehua Qin. 2019. A method of Chinese tourism named entity recognition based on BBLC Model. Em *IEEE SmartWorld: Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation*, 1722–1727.

Proposta recebida em Julho 2023 e aceite para publicação em Novembro 2023.

Transferência de estilo textual arbitrário em português

Arbitrary Portuguese text style transfer

Pablo Botton da Costa
Universidade de São Paulo (EACH-USP)

Ivandré Paraboni
Universidade de São Paulo (EACH-USP)

Resumo

Na Geração automática de língua natural, modelos de transferência de estilo textual arbitrário objetivam a reescrita de um texto usando qualquer novo conjunto de características estilísticas desejado. Em se tratando do idioma português, entretanto, observa-se que os recursos linguístico-computacionais necessários para o desenvolvimento de modelos deste tipo ainda são consideravelmente escassos em comparação à língua inglesa. Assim, como um primeiro passo em direção ao desenvolvimento de métodos avançados deste tipo, o presente trabalho investiga a questão da transferência de estilo textual arbitrário com o uso de paráfrases em português, combinadas ao uso de modelos neurais construídos a partir de arquiteturas do tipo *sequência-para-sequência* e por refinamento de grandes modelos de língua. Além dos modelos de reescrita textuais propriamente ditos, o estudo apresenta também recursos inéditos para a tarefa na forma de um córpus de paráfrases e de um modelo de *embeddings* validado nas tarefas de similaridade e simplificação sentencial, com resultados comparáveis ao estado da arte.

Palavras chave

geração de língua natural, transferência de estilo arbitrário, paráfrases, sequência-para-sequência, grandes modelos de língua

Abstract

In Automatic Natural Language Generation, arbitrary style transfer models aim to rewrite a text using any desired new set of stylistic features. In the case of the Portuguese language, however, we notice that the resources required for the development of models of this type are still considerably scarce compared to those dedicated to the English language. Thus, as a first step towards the development of advanced methods of this kind, the present work investigates the issue of arbitrary style transfer with the aid of paraphrases in Portuguese, combined with the use of neural models built from sequence-to-sequence architectures and by refining a number of large language models. In addition to the textual rewriting models themselves, the study also presents novel resources for the task in the form of a corpus of paraphrases and a model of embeddings validated in both sentence similarity and simplification tasks, with results comparable to the state of the art.

Keywords

natural language generation, arbitrary style transfer, paraphrases, sequence-to-sequence, large language models

1. Introdução

O uso de métodos de aprendizado neural tem se tornado prevalente em diversas áreas do Processamento de Língua Natural (PLN), incluindo a Geração de Língua Natural (GLN) (Gatt & Krahmer, 2018; Naseem et al., 2021; Dong et al., 2022). Dentre múltiplas aplicações atuais deste tipo, destacamos no presente estudo a tarefa computacional de *transferência de estilo* (Jin et al., 2022; Hu et al., 2022), que consiste em reescrever um texto sem alterar seu conteúdo semântico, mas em um estilo[1] diferente do original. Modelos de transferência de estilo têm sido aplicados em cenários específicos como a transferência de sentimentos (Hu et al., 2017; Shen et al., 2017; Li et al., 2018; Luo et al., 2019), formalidade (Xu et al., 2012; Wang et al., 2019) e outros, de modo geral fazendo uso de um córpus alinhado de pares de frases em um estilo-origem e um estilo-alvo de interesse.

Mais recentemente, a tarefa de transferência de estilo passou a ser investigada também em cenários onde os estilos-alvo da geração são *arbitrários* (Krishna et al., 2020; Reif et al., 2022; Suzgun et al., 2022), ou seja, utilizando-se de modelos que permitem a reescrita de textos com qualquer novo conjunto de características estilísticas desejado. Abordagens deste tipo dis-

[1] Considerando-se aqui a definição de estilo proposta em Jin et al. (2022), que engloba qualquer atributo que varia de um texto origem para um texto alvo, e não apenas no sentido estritamente linguístico do termo.

DOI: 10.21814/lm.15.2.410

pensam o uso de córpus previamente alinhados, que são substituídos pelo uso de algum tipo de alinhamento parcial pela geração do tipo *sequência-para-sequência* (Riley et al., 2021), pelo refinamento de grandes modelos de língua pré-treinados (LLMs), ou por meio de paráfrases (Krishna et al., 2020). Esta última opção será também o foco do presente trabalho.

De forma mais específica, o presente estudo aborda a tarefa computacional de geração textual de estilo baseada na noção de *quase-paráfrases*, definidas em (Bhagat & Hovy, 2013) como um par de frases ou sentenças que transmitem aproximadamente o mesmo significado usando palavras diferentes. As sentenças 1 e 2 a seguir, que podem ser vistas como sendo duas formas distintas (e.g., mais ou menos formal) de expressar uma mesma ideia geral, são exemplos de quase-paráfrases.

1 Foi informado pela *instituição* que seus ônibus *teriam capacidade para* 40 alunos cada.

2 A *escola falou* que seus ônibus *acomodam* 40 alunos cada.

Considerando-se que um conjunto de exemplos em um estilo fonte (como 1 acima) e outro estilo alvo (como 2) esteja disponível em forma de um córpus alinhado de paráfrases, métodos de aprendizado neural podem ser aplicados para identificar as relações de paráfrase entre seus termos, que podem então ser utilizadas na reescrita de uma sentença qualquer mesmo que esta não ocorra no conjunto de treino. Em outras palavras, modelos deste tipo, que efetivamente implementam a tarefa computacional de transferência de estilo, são capazes de reescrever um texto arbitrário qualquer no estilo fonte para o estilo alvo, como no caso da reescrita de um texto formal como o exemplo 1 acima em uma versão mais informal, como o exemplo 2.

A transferência de estilo baseada em paráfrases é de grande interesse científico, e possui uma ampla gama de aplicações possíveis (Jin et al., 2022). No entanto, no caso específico da língua portuguesa, observa-se que tanto os métodos de Geração de Língua Natural como os recursos linguístico-computacionais necessários para esse fim ainda são consideravelmente escassos em comparação com o estado da arte disponível para a língua inglesa. Assim, como um primeiro passo em direção ao desenvolvimento de métodos avançados deste tipo, o presente trabalho objetiva investigar a questão da transferência de estilo textual arbitrário por meio de paráfrases em português, apresentando recursos inéditos para a tarefa e métodos de transferência de estilo treinados e avaliados com base nestes recursos.

As principais contribuições previstas neste estudo são as seguintes:

- Modelos computacionais de transferência de estilo textual arbitrário em português construídos a partir de arquiteturas do tipo *sequência-para-sequência* e por refinamento de grandes modelos de língua.
- Resultados de avaliação dos modelos propostos com uso de métricas automáticas e avaliação humana.
- Córpus de paráfrases alinhadas em nível sentencial.
- Modelo de *embeddings* de paráfrases para transferência de estilo.
- Resultados de avaliação destes recursos nas tarefas de simplificação e similaridade sentenciais, comparáveis ao estado da arte.

O restante deste artigo é organizado da seguinte forma. A seção 2 sumariza a pesquisa recente em transferência de estilo textual arbitrário para o idioma inglês. A seção 3 descreve a construção do córpus de paráfrases a ser utilizado como base para os modelos propostos. A seção 4 descreve a construção e validação de um modelo de *embeddings* de paráfrases tomando como exemplo as tarefas de simplificação e similaridade sentencial. Com base nestes recursos, a seção 5 apresenta nossa abordagem principal de reescrita sentencial para transferência de estilo. Finalmente, a seção 6 apresenta as conclusões do presente estudo e opções de trabalhos futuros.

2. Trabalhos relacionados

A tarefa de transferência de estilo textual arbitrário ainda é relativamente pouco explorada na pesquisa em GLN. A seguir revisamos brevemente os estudos em Krishna et al. (2020); Riley et al. (2021); Reif et al. (2022) e outros trabalhos a eles relacionados, que propõem a utilização de grandes modelos de língua como forma de se obter e gerar dados sintéticos em cenários em que haja poucos recursos disponíveis para a tarefa. Um levantamento mais detalhado da área é apresentado em Jin et al. (2022).

O trabalho em Krishna et al. (2020) propõe um *framework* de geração do tipo *texto-para-texto* dito universal, ou seja, para qualquer estilo, sem a necessidade prévia de dados alinhados. O *framework* de geração é composto de

duas etapas. A primeira é responsável pela conversão do texto de entrada em uma paráfrase, desempenhada por um modelo previamente treinado em um conjunto aprimorado de pares de paráfrases do córpus *PARANMT-50M* (Wieting et al., 2017). Os pares candidatos são então selecionados a partir do córpus com base em filtros com o objetivo de maximizar sua diversidade lexical e complexidade, e usados para o ajuste de pesos do modelo GPT-2. A segunda etapa consiste de um modelo de geração do tipo *sequência-para-sequência* treinado para inverter paráfrases em onze estilos. Por exemplo, um texto de entrada como "*I'd jump in there, no doubt*", reescrito em estilo de paráfrase durante a etapa anterior, seria convertido para o estilo original como "*No lie... I would jump in*". Cada *framework* é treinado individualmente, e a transferência de estilo é realizada por meio da substituição do módulo de inversão de paráfrases pelo módulo de inversão de paráfrases do estilo alvo desejado. Por exemplo, dado um texto de entrada e o objetivo de reescrevê-lo no estilo da língua inglesa moderna, o modelo gera a paráfrase dessa entrada, e a reescreve no estilo-alvo desejado com o uso respectivo modelo de inversão de paráfrase.

O modelo proposto, chamado STRAP, é avaliado com uso de cinco métricas automáticas e humanas a partir dos córpus supervisionados *Shakespeare* (Xu et al., 2012) e *Formality* (Rao & Tetreault, 2018). As métricas de avaliação automática utilizadas foram (1) a força da transferência de estilo (aferida por um classificador de estilos treinado a partir de textos dos córpus *Shakespeare* e *Formality*), (2) a semelhança de paráfrases (calculada como a semelhança de cosseno da representação distribuída dos pares de textos a partir dos pesos apresentados em Wieting et al. (2021)), (3) a fluência do texto gerado (computada por um modelo de classificação treinado a partir do córpus de aceitabilidade gramatical COLA, como em Warstadt et al. (2019)), (4) a média geométrica entre as métricas de força da transferência, similaridade de paráfrases e fluência dos textos, e (5) o produto entre as métricas força da transferência, similaridade de paráfrases e a fluência dos textos dividido pelo tamanho da sentença alvo. Para as métricas de avaliação humana, foi proposto a um grupo de juízes medir (1) a adequação semântica, (2) a dissimilaridade lexical, e (3) a semelhança estilística dos textos gerados pelos modelos, conforme os critérios do *crowdsourcing*.

O trabalho em Riley et al. (2021) propõe a adaptação do modelo T5 (Raffel et al., 2020) como um extrator de características para a tarefa de transferência de estilo, seguindo a abordagem em Biber & Conrad (2019) de treinamento de uma rede neural do tipo *denoising auto-encoder* condicionado a um vetor de estilo. Dado que o vetor de estilo original é desconhecido, a abordagem constrói de forma conjunta um modelo neural para induzir a representação de estilo do texto. Por exemplo, dado um texto origem e o objetivo de reescrevê-lo no estilo contrário ao sentimento original da frase, o modelo gera dois novos exemplos: o primeiro tem estilo similar ao da entrada, mas com palavras diferentes, e o segundo é um texto aleatório do conjunto de treinamento que pertence ao estilo-alvo do objetivo de geração. Estas três entradas – os dois exemplos e o texto origem — são então submetidos ao modelo para que sejam combinados de tal forma a reescrevê-los no estilo-alvo desejado.

O modelo obtido, chamado *TextSETTR*, é avaliado a partir da combinação entre métricas humanas e automáticas em relação ao conjunto de testes. Para a métrica de avaliação automática, optou-se pela utilização das métricas BLEU e força da transferência de estilo aferida por meio de um classificador treinado para identificar estilos. O modelo foi avaliado em três tarefas de transferência de estilo usando os córpus *Amazon Reviews*, *Shakespeare* e o córpus de textos bíblicos em Carlson et al. (2018). Dentre outros resultados, observa-se que modelo *TextSETTR* apresenta uma melhoria discreta em relação ao trabalho em Biber & Conrad (2019).

Finalmente, o trabalho em Reif et al. (2022) propõe o uso da técnica de engenharia de *prompts* em combinação com métodos de destilação de conhecimento a partir de grandes modelos de língua da família GPT (Brown et al., 2020) para a tarefa de transferência de estilo. Por exemplo, dado um texto de entrada e o objetivo de reescrevê-lo no estilo de poesia, e.g., parnasiana, o modelo é instruído com o *prompt* '*transforme o texto a seguir em estilo de poesia parnasiana* $\rightarrow x'$, resultando no texto x' reescrito no estilo-alvo desejado. O trabalho propõe a construção de cinco modelos: os *baselines GPT-3-ada*, *GPT-3-curie* e *GPT-3-davinci*, que não utilizam *prompts*; um modelo que utiliza uma rede neural do tipo *transformers Lambda* treinada em um córpus composto por 1.95B documentos de domínio público, e um modelo chamado *lambdaFinne* que utiliza a mesma arquitetura base do anterior, porém com os pesos ajustados em um córpus com curadoria de alta qualidade para o domínio conversacional.

A avaliação destes modelos foi realizada de duas formas. Uma considerou estilos atípicos ou sem padrão definido, correspondendo aos ajustes

mais frequentes feitos por usuários de uma ferramenta inteligente de edição de textos, e a outra considerou os estilos pré-definidos de sentimento (presente no córpus Yelp em Zhang et al. (2015)) e formalidade (do córpus GYAFC em Rao & Tetreault (2018)).

Os estudos de transferência de estilo aqui discutidos são todos dedicados ao idioma inglês, e fazem uso de recursos linguístico-computacionais ainda escassos para o idioma português, como córpus alinhados, *embeddings* de paráfrases e grandes modelos de língua. Estas observações foram levadas em conta na construção de recursos básicos deste tipo para o português, e no método de reescrita sentencial baseada em paráfrases a serem abordados pelo presente estudo.

3. Córpus de paráfrases PTPARANMT

O uso de paráfrases tem se destacado como um método robusto para reescrita de texto em um estilo-alvo diferente do seu estilo original (Krishna et al., 2020, 2022). Um estudo deste tipo requer, entretanto, um córpus paralelo contemplando textos alinhados em nível de sentenças, ou seja, um conjunto de textos origem a ser parafraseado, e um segundo conjunto de textos com o mesmo significado, porém com características lexicais e sintáticas modificadas. Em virtude da dificuldade de obtenção de um recurso linguístico deste tipo em português, e com qualidade e volumes adequados para a tarefa de transferência de estilo, optou-se por criar um novo córpus de paráfrases aos moldes de Wieting et al. (2017), aqui denominado *PTPARANMT*. A construção e avaliação deste conjunto de dados é o foco desta seção.

O córpus *PTPARANMT* foi criado a partir da tradução reversa, ou seja, traduzindo-se textos de um idioma para outro, e novamente para o idioma original como forma de provocar modificações léxicas e estruturais com pouca ou nenhuma perda de significado. Por exemplo, o texto "eu te amo" poderia ser traduzido para o inglês como "i love you," e então traduzido no sentido reverso para "te amo." Técnicas de tradução reversa têm sido aplicadas com sucesso em tarefas como tradução automática (Hoang et al., 2018), transferência de estilo (Krishna et al., 2020, 2022) e sumarização automática (Beddiar et al., 2021), além da própria construção de córpus a partir de recursos já disponíveis (Gonçalo Oliveira & Alves, 2021).

Foram tomados por base três córpus existentes (*Europarl*, *ParaCrawl* e *Tapaco*) utilizando-se das porções alinhadas Português-Inglês de cada um. A escolha deste par de idiomas foi motivada pela alta incidência de paráfrases de qualidade obtidas a partir destes córpus para fins de tradução automática em Wieting et al. (2017). O córpus *Europarl*, já utilizado em Wieting & Gimpel (2018) para geração de paráfrases, é constituído de transcrições das seções públicas do parlamento europeu em 21 idiomas, e possui 1.960.407 pares de sentenças alinhadas Português-Inglês. O córpus *ParaCrawl*, já utilizado em Bañón et al. (2020) para a tarefa de tradução automática, é constituído de textos provenientes de páginas da web em 42 idiomas, e possui 84.921.510 pares de sentenças alinhadas Português-Inglês. Por se tratar de um córpus extenso, foi realizada uma amostragem de 19 milhões de pares deste córpus. Finalmente, o córpus *Tapaco*, já utilizado em Shliazhko et al. (2022) na tarefa de identificação de paráfrases, é constituído de textos da base *Tatoeba*, construída por *crowdsourcing* em 73 idiomas, e possui 110.000 pares de sentenças alinhadas Português-Inglês.

A construção do córpus *PTPARANMT* pode ser dividida em duas etapas. Na primeira, foi gerada uma base de pares de textos alinhados em nível sentencial com uso de tradução reversa a partir dos três córpus de entrada. Na segunda, o conjunto de textos foi filtrado com base em um limite de confiança estimado por meio de avaliação humana, resultando no córpus final. Estas duas etapas são discutidas individualmente a seguir.

3.1. Base de textos alinhados

O córpus *PTPARANMT* consiste de dois conjuntos de textos em português alinhados em nível sentencial, aqui denominados Origem e Alvo. O conjunto Origem foi obtido a partir da tradução do original inglês para o português, e o conjunto Alvo foi obtido pela tradução dos textos originais em português para o inglês, e posteriormente traduzidos de volta para o português.

Para as etapas de tradução e tradução reversa, foi utilizado o *framework OPUS-MT* (Tiedemann & Thottingal, 2020). A tradução do conjunto Origem usou a versão do *framework* en-ROMANCE, e a etapa de tradução reversa usou as versões ROMANCE-en e en-ROMANCE para a tradução para o inglês e para o português, respectivamente.

A tradução e tradução reversa dos conjuntos Origem e Alvo geraram 21 milhões de pares de textos alinhados em português. A Tabela 1 sumariza as estatísticas descritivas deste conjunto de dados a partir da amostra dos pares presentes em cada uma das três origens textuais consideradas. Estas estatísticas descrevem o tamanho médio das sentenças, a métrica de paráfrases *para-score* calculada pela distância média de cosseno entre as *embeddings* sentenciais de forma relativa aos pares de paráfrases, cf. Wieting et al. (2017), o grau de sobreposição média de bigramas e trigramas entre pares de sentenças, escores BLEU (Post, 2018), distância de edição (DE) e quantidade total de sentenças.

Conforme ilustrado na Tabela 1, o córpus *Europarl* tem a melhor média para a métrica de similaridade de paráfrases. O córpus *Tapaco*, por outro lado, apresenta em média as menores frases e as menos diversas.

Dado que o conjunto de dados inclui textos provenientes da Internet, foram excluídas as sentenças com confiança $\leq 0,97$ para o classificador,[2] que determina se a linguagem resultante do processo é ou não português. Além disso, foram removidas as sentenças com menos de três palavras, ou que apresentavam erros como palavras repetidas ou reticências. A Tabela 2 apresenta exemplos de textos do conjunto Origem e Alvo apresentando relação de paráfrase do par de texto e a confiança do classificador da língua portuguesa em relação ao texto Origem.

3.2. Filtragem

Após a construção do conjunto de textos inicial, foi conduzida uma avaliação humana com o objetivo de reduzi-lo ao subconjunto de paráfrases de maior confiança, ou seja, eliminando-se tanto quanto possível o ruído existente. Para este fim, oito falantes nativos da língua portuguesa foram solicitados a avaliar 63 pares de paráfrases provenientes de uma amostragem obtida a partir de diferentes faixas da métrica *para-score*. Entretanto, dado que os textos podem conter diversos tipos de imperfeição (e.g., decorrente do córpus de origem ou do processo de tradução, dentre outras possibilidades), foi solicitado aos participantes que avaliassem cada um dos pares de frases quanto aos possíveis *erros* de paráfrase (ou seja, o quanto uma frase pode ser considerada como sendo uma paráfrase da sua contrapartida) e fluência.

[2]Language Detection Library for Java, `http://code.google.com/p/language-detection/`

A avaliação de erros usou uma versão adaptada das instruções apresentadas em Agirre et al. (2012) na qual os rótulos ('alto', 'médio' e 'baixo') originalmente propostos para avaliar a força da paráfrase ou fluência do texto foram usados de forma invertida para representar o grau de erro segundo estes dois critérios.

Os três graus de erros de paráfrase considerados foram os seguintes:

- erro *baixo*: as sentenças devem ter o mesmo significado, mas alguns detalhes sem importância podem diferir.
- erro *médio*: as sentenças devem ser aproximadamente equivalentes, com algumas informações importantes faltando ou diferem um pouco.
- erro *alto*: as sentenças não são equivalentes, mesmo que compartilhando pequenos detalhes.

De forma análoga, a avaliação de fluência considerou os três graus de erro a seguir:

- erro *baixo*: o texto não deve conter erros gramaticais.
- erro *médio*: o texto deve possuir um ou dois erros gramaticais.
- erro *alto*: o texto deve possuir mais de dois erros gramaticais, ou não soa natural em português.

Foi alcançada uma concordância entre anotadores de $0,89$ de acordo com o índice kappa de Cohen (Landis & Koch, 1977). A Tabela 3 apresenta os resultados da avaliação humana agrupados em intervalos de *para-score*, considerando-se as medidas da média da sobreposição de trigrama, o número de vezes que cada texto Origem ou Alvo recebeu um escore de fluência 0, 1 ou 2 (denominados Fluência.O e Fluência.A, respectivamente), e o número de vezes que cada texto Origem ou Alvo recebeu um escore de paráfrase 0, 1 ou 2.

Conforme pode ser observado na Tabela 3, os pares ruidosos estão, em sua maioria, confinados ao primeiro intervalo $(0,24, 0,631)$. Este intervalo compreende 19,23% de todos os pares que possuem uma forte relação de paráfrase. Além disso, nos dois intervalos mais altos, 84% dos pares possuem uma relação forte de paráfrases. Nos intervalos baixos, por outro lado, uma inspeção manual dos dados originais revelou principalmente erros de alinhamento em nível de sentença, o que também foi reportado no trabalho para a língua inglesa em Wieting & Gimpel

Origem	Tamanho	para-score	sobrepos.	BLEU	DE	Sentenças
Europarl	25,08	0,80	0,44	71,65	61,26	1,9 M
ParaCrawl	14,45	0,75	0,34	54,03	47,16	19,0 M
Tapaco	5,89	0,72	0,47	63,59	10,86	0,110 M

Tabela 1: Estatísticas de 100.000 amostras de paráfrases de cada origem textual utilizada.

para-score	Confiança	Origem	Alvo
0,19	0,98	*com certeza.*	*tem toda a razão.*
0,36	0,94	*— ela é colombiana.*	*ela é irlandesa.*
0,51	1,00	*embora, como você tenha visto, o temido erro do milênio não tenha se materializado, ainda assim as pessoas de vários países sofreram uma série de desastres naturais que realmente foram terríveis.*	*como puderam constatar, o grande "bug do ano 2000" não aconteceu. em contrapartida, os cidadãos de alguns dos nossos países foram vítimas de catástrofes naturais verdadeiramente terríveis.*
0,70	1,00	*senhora presidente, gostaria de saber se haverá uma mensagem clara do parlamento esta semana sobre o nosso descontentamento sobre a decisão de hoje se recusar a renovar o embargo de armas sobre a indonésia, considerando que a grande maioria neste parlamento aprovou o embargo de armas na indonésia no passado?*	*senhora presidente, gostaria de saber se esta semana o parlamento terá oportunidade de manifestar a sua inequívoca posição de descontentamento face à decisão, hoje tomada, de não renovar o embargo de armas destinadas à indonésia, tendo em atenção que a grande maioria da assembleia apoiou o referido embargo quando este foi decretado.*
0,97	0,98	*ele reuniu diferentes exemplos.*	*reuniu diferentes exemplos.*
1,00	1,00	*é tão fácil. . .*	*é tão fácil. . .*

Tabela 2: Pares de amostras ordenados por grau de relação de paráfrase (para-score) e grau de confiança da sentença de Origem em relação à língua portuguesa.

(2018). Para a métrica de fluência, aproximadamente 82,53% dos textos oriundos do processo de tradução reversa são fluentes.

O subconjunto de pares de paráfrases dos dois níveis superiores será utilizado nos experimentos descritos nas próximas seções. Estatísticas descritivas são apresentadas na Tabela 4.

3.3. Exemplos de paráfrases obtidas

Como forma de exemplificar a qualidade dos pares de paráfrases presentes no córpus *PTPARANMT*, a Tabela 5 apresenta exemplos de textos e suas paráfrases selecionados aleatoriamente. Para facilitar a ilustração, os exemplos são agrupados informalmente em três categorias de erro (baixa, média e alta) conforme seu valor de *para-score*. De forma análoga, a Tabela 6 apresenta exemplos de textos do conjunto Origem apresentando erro de fluência baixo, médio ou alto. Em ambos os casos, observa-se que a qualidade dos textos utilizados no processo de construção do córpus (seções 3.1 e 3.2) é variável, o que se reflete naturalmente na qualidade das paráfrases obtidas.

4. Indução de embeddings sentenciais

O córpus de paráfrases *PTPARANMT* foi utilizado para indução de *embeddings* sentenciais para uso nos experimentos de transferência de estilo a serem relatados na Seção 5, privilegiando-se para este fim os pares de sentença que permitissem maximizar a variedade lexical com preservação da semântica original do texto. Assim como no estudo em Wieting & Gimpel (2018) para a língua inglesa, optamos por selecionar os pares de sentenças cujo *para-score* fosse $\geq 0,4$, e foram excluídos os pares cuja sobreposição de trigramas fosse $\leq 0,7$. A aplicação destes filtros resultou em um subconjunto de 13 milhões de pares de paráfrases a ser utilizado como conjunto de treino para as *embeddings*.

As *embeddings* foram geradas a partir de pares de córpus conforme a metodologia em Wieting et al. (2021). Essa metodologia propõe uma arquitetura neural composta por uma camada densa projetada para um vetor que combina, através da média, todas as representações de todos os trigramas possíveis para uma sentença s. O objetivo da modelagem é, a partir de

Agrupamentos por *para-score*	# pares	Sobreposição Trigramas	Fluência.O			Fluência.A			Paráfrase		
			0	1	2	0	1	2	0	1	2
$(0,24, 0,631)$	13	$0,22 \pm 0,12$	10	11	05	16	05	05	05	13	08
$(0,631, 0,752)$	12	$0,38 \pm 0,10$	12	05	07	16	04	04	16	05	03
$(0,752, 0,841)$	13	$0,47 \pm 0,12$	08	10	08	14	09	03	22	02	02
$(0,841, 0,913)$	12	$0,47 \pm 0,17$	16	07	01	12	10	02	20	04	00
$(0,913, 1,0)$	13	$0,63 \pm 0,16$	12	09	05	14	08	04	22	04	00

Tabela 3: Avaliação humana dos textos coletados.

	Origem	Alvo
Pares	21.355.451	18.650.340
Palavras	1.349.479	815.791
Tam. sentenças	6,14	6,15

Tabela 4: Estatísticas descritivas do córpus.

uma sequência de palavras (s), extrair um vetor denso de *embeddings* de tamanho k minimizando-se a função de similaridade entre pares similares e, para casos negativos, maximizando-se a função de recompensa.

Para os hiper-parâmetros, seguimos as mesmas configurações usadas em Wieting et al. (2021), com *batch* de tamanho 128, *margin* σ de 0,4, e a taxa de *annealing* em 150. Um *tokenizador* do tipo *SentencePiece* (Kudo & Richardson, 2018) foi pré-treinado usando-se o córpus com vocabulário máximo de 50.000 *tokens*. O otimizador utilizado foi o *adam* (Kingma & Ba, 2015) com uma taxa de aprendizado inicial em 0,001, e iterando-se o modelo por um total de 25 épocas. Como forma de otimizar o processo de treinamento, optou-se pela utilização do método *mega-batch* com tamanho 100, e *dropout* inicial para a camada intermediária de 0,0.

A partir de uma sentença s_i, o treinamento propriamente dito consistiu em selecionar como alvo negativo uma sentença aleatória t'_i que não fosse uma paráfrase t_i. Esta seleção é feita com base nas sentenças do conjunto Alvo do córpus, em todos os casos selecionando-se o exemplo negativo com menor similaridade de cosseno em relação à sentença. Esta estratégia objetivou obter pares positivos que fossem mais similares do que os pares de sentença negativos com a seguinte margem α:

$$\min_{(\sigma_{origem}, \sigma_{alvo})} \sum \alpha - \cos_\sigma(s_i, t_i) + \cos_\sigma(s_i, t'_i)$$

O modelo assim definido foi treinado com uso do algoritmo de *back-propagation*. Como forma de ilustrar a qualidade das *embeddings* geradas, foram conduzidos dois experimentos de avaliação intrínseca envolvendo tarefas tradicionais de PLN que guardam certa afinidade com a detecção de paráfrases. Estes experimentos, enfocando as tarefas de simplificação, e similaridade sentencial, são descritos individualmente nas seções a seguir.

4.1. Simplificação sentencial

A tarefa de simplificação sentencial considerada para avaliação das *embeddings PTPARANMT* consistiu em criar versões de menor complexidade lexical e sintática de um texto de entrada nos moldes definidos pelos córpus PorSimplesSent2 e PorSimplesSent3 descritos em Leal et al. (2018).

Foram desenvolvidos três modelos de simplificação sentencial do tipo *sequência-para-sequência* com arquitetura neural do tipo *transformer*, aqui denominados S2S+PTPARANMT, *S2S+Glove* e *S2S+Random*. No modelo *S2S+PTPARANMT*, optamos por utilizar um esquema de treinamento *sequência-para-sequência* com *transfer-learning* usando os pesos das *embeddings* induzido a partir do córpus *PTPARANMT*. No modelo *S2S+Glove* usamos *embeddings* do tipo GloVe (Pennington et al., 2014). No modelo *S2S+Random*, optamos por utilizar um esquema de inicialização de pesos aleatória.

Como hiper-parâmetros dos modelos, definimos um *batch* de tamanho 32, e usamos o otimizador *adam* com uma taxa de aprendizado inicial de 0,001 em um total de 20000 épocas com um intervalo de tolerância de 10 épocas para paradas preemptivas. Optamos por usar a métrica de avaliação automática BLEU por se tratar de um problema de geração de texto. As *embeddings* GloVe utilizadas são disponibilizadas em Hartmann et al. (2017).

Como pré-processamento das versões 2 e 3 do córpus PorSimplesSent, utilizamos um *tokenizador* do tipo *SentencePiece* (Kudo & Richardson, 2018) previamente treinado com base nos dados do córpus *PTPARANMT*. A divisão do córpus foi de 80% para treino, 10% para testes e 10% para o conjunto de treinamento para a validação.

Erro	Origem	Alvo
Baixo	*graças a esse apoio, podemos corrigir o ângulo da câmera.*	*graças a este suporte, podemos corrigir o ângulo da câmera.*
Médio	*eu vou revelar a resposta para você e, em seguida, dar-lhe uma meditação guiada para incorporar os conceitos em sua mente subconsciente para que você seja naturalmente mais feliz sozinho ou para você aprender a ser sozinho.*	*pode ser realmente feito? im irá revelar a resposta para você e, em seguida, dar-lhe uma meditação guiada para incorporar os conceitos na sua mente subconsciente, de modo que você naturalmente seja ainda mais feliz sozinho.*
Alto	*uma vez por turno: você pode desligar o material de 1 xyz desta carta, em seguida, alvo de 1 monstro que o seu adversário controla; equipa- o para esta carta.*	*uma vez por turno, você pode selecionar uma face voltada para cima do monstro synchro seu oponente controla, e equipá-la a este cartão.*

Tabela 5: Pares de amostras ordenados por grau de erro de paráfrase.

Erro	Origem
Baixo	*infundindo os clássicos europeus com toque lúdico e moderno os nossos programas experimentais que definem marca dão vida ao le méridien e cumprem nossa promessa de desbloquear destino através de experiências criativas e culturais para os hóspedes.*
Médio	*no ano seguinte huam tchao organizou insurreição em apoio vam sientchi.*
Alto	*andrea dovizioso repsol honda 2010 motogp couro terno.*

Tabela 6: Amostras do conjunto Origem ordenadas por grau de erro de fluência.

Os modelos foram treinados na porção de testes dos conjuntos de dados propostos. A Tabela 7 sumariza as configurações empregadas e seus resultados com base nas versões 2 e 3 do córpus (v2-BLEU e v3-BLEU).

Os resultados da Tabela 7 sugerem que o modelo *S2S+PTPARANMT* obteve os melhores resultados gerais. Além disso, cabe destacar a importância do uso de pesos pré-treinados, ilustrada pelo desempenho inferior do modelo *S2S+Random* em relação aos dois primeiros.

4.2. Similaridade sentencial

Como um segundo cenário de avaliação das *embeddings PTPARANMT*, consideramos a tarefa de estimativa de similaridade sentencial. Essa tarefa consistiu em decidir - em uma escala de 0 a 5 - quão próximas duas sentenças são entre si tomando-se por base o *benchmark ASSIN* 2 (Real et al., 2020). Com este propósito, foi conduzido um experimento comparando-se dois modelos refinados a partir dos pesos das *embeddings* geradas, aqui denominados *PTPARANMT* e *Siamese+PTPARANMT*, e dois sistemas de *baseline* do tipo *transformer*, aqui denominados *para-multi-MiniLM-L12-v2* e *dpr-ctx-enc-bert-base-multi*.

O modelo *PTPARANMT* utiliza a distância cosseno entre as *embeddings* dos pares de sentença, enquanto *Siamese+PTPARANMT* utiliza uma rede recorrente do tipo LSTM para a sentença Origem e outra para a Alvo, tal que ambas são otimizadas de modo a minimizar a distância cosseno entre os pares candidatos. O modelo de *baseline para-multi-MiniLM-L12-v2* usa uma arquitetura *transformer* similar ao apresentado em Reimers & Gurevych (2019), utilizando os pesos multilíngues do modelo original em um formato de *transfer-learning*. O *baseline dpr-ctx-enc-bert-base-multi*, por outro lado, adota o princípio de passagem densa de representação (do inglês *Dense Passage Retrieval*), em um formato similar ao apresentando em Karpukhin et al. (2020), também utilizando o peso multilíngue do modelo original em um formato de *transfer-learning*.

Como hiper-parâmetros dos modelos, definimos um *batch* de tamanho 32, otimização *adam* para uma taxa de aprendizado inicial de 0,001 em um total de 1000 épocas com um intervalo de tolerância de 10 épocas para paradas preemptivas.

Para pré-processamento do córpus *ASSIN* 2, utilizamos um *tokenizador* do tipo *SentencePiece* (Kudo & Richardson, 2018), que foi treinado previamente nos dados do córpus *PTPARANMT*. O conjunto de dados *ASSIN* 2 contém 10.000 pares de frases em português brasileiro, sendo 6.500 pares para treinamento, 500 para validação e 3.000 para teste. As configurações empregadas e seus resultados — representados pelas métricas *Mean*

Modelo	Tam.	Neurônios	Camadas	Pré-treino?	v2-BLEU	v3-BLEU
S2S+PTPARANMT	1024	512	12	Sim	79,8	52,0
S2S+Glove	300	512	12	Sim	70,2	43,9
S2S+Random	300	512	12	Não	60,9	30,9

Tabela 7: Modelos e resultados de simplificação sentencial.

Square Error (MSE) e correlação de Pearson — são sumarizados na Tabela 8, com os melhores resultados de cada métrica em destaque.

A Tabela 8 mostra que o modelo *Siamese +PTPARANMT* obteve os melhores resultados gerais para ambas as métricas de avaliação. Com base neste resultado, o modelo *Siamese+PTPARANMT* foi então comparado também com os sistemas participantes da iniciativa *ASSIN* 2. Os resultados desta avaliação são sumarizados na Tabela 9, novamente com os melhores resultados de cada métrica em destaque.

Os resultados da Tabela 9 indicam que o modelo *Siamese+PTPARANMT* supera as alternativas tanto no que diz respeito à métrica MSE (i.e., comparado ao sistema Stilingue em Fonseca & Alvarenga (2019)), quanto no que diz respeito à correlação de Pearson (comparado ao sistema IPR em Rodrigues et al. (2019b)).

5. Reescrita sentencial baseada em paráfrases

Os resultados do uso do córpus e *embeddings PTPARANMT* nas tarefas de simplificação (Seção 4.1) e similaridade sentencial (Seção 4.2) sugerem que estes recursos podem ser utilizados na aplicação-fim do presente estudo, ou seja, à tarefa de transferência de estilo arbitrário baseada em paráfrases. Para investigar esta possibilidade, foi conduzido um experimento deste tipo utilizando modelos neurais do tipo *sequência-para-sequência* para reescrita sentencial em um estilo-alvo de interesse. A escolha dessa arquitetura foi motivada pela sua relativa simplicidade de implementação, e pelo bom desempenho geral observado em tarefas de geração de texto (Goodfellow et al., 2016; Goldberg, 2016; Gatt & Krahmer, 2018).

O experimento realizado consistiu em desenvolver e comparar duas estratégias de reescrita sentencial — aqui denominadas *S2S* e *paraPTT5* — com três modelos de *baseline*, aqui denominados *REF*, *Cópia* e *Ingênuo*. Estas cinco configurações são sumarizadas na Tabela 10 e detalhadas a seguir.

O modelo *S2S* consiste de uma arquitetura neural do tipo *sequência-para-sequência* com atenção do tipo geral utilizando o *framework* de geração de texto *openNMT* com as mesmas configurações descritas em Bahdanau et al. (2014), e com inicialização de pesos no formato xavier (Glorot & Bengio, 2010). O modelo *paraPTT5* consiste de uma arquitetura neural de pesos ajustados a partir do modelo de língua pré-treinado *PTT5* (Carmo et al., 2020). O ajuste dos pesos originalmente fornecidos pelo modelo fez uso da biblioteca *Transformers Hugging Face* (Wolf et al., 2020). Em ambos os casos, termos desconhecidos do vocabulário do modelo foram representados por *tokens* artificiais *UNKNOWN*.

Como sistemas de *baseline* para o experimento, consideramos três modelos que imitam transformações simples sobre a entrada textual: o modelo *REF* apenas reproduz o texto alvo como saída, e representa assim o limite de desempenho máximo possível para a tarefa; *Cópia* é uma simulação de um modelo de geração que simplesmente repete a entrada como saída; e *Ingênuo* é um modelo que gera como saída uma seleção aleatória das palavras de entrada, com probabilidade $p = 0,5$, e concatena a elas um texto também aleatório do conjunto Alvo.

5.1. Conjunto de dados

A partir da versão original do córpus *PTPARANMT* (descrita na seção 3), foram aplicados filtros sistemáticos com o objetivo de produzir um conjunto de dados de modo que as características de diversidade lexical e sintática fossem maximizadas. Esse tipo de técnica tem sido aplicada com resultados positivos em tarefas para a língua inglesa como reescrita de paráfrases (Wieting et al., 2021) e transferência de estilo (Krishna et al., 2020).

Para filtrar o córpus *PTPARANMT*, utilizamos as mesmas duas métricas originalmente propostas em Krishna et al. (2020). A métrica *sobreposição* foi usada para medir a diversidade lexical dos textos a partir da contagem de sobreposição de trigramas de co-ocorrência entre os pares, e a métrica *para-score* foi usada para medir a similaridade de paráfrase entre dois textos.

Modelo	Tam.	Neurônios	Camadas	MSE	Pearson
PTPARANMT	1024	1024	1	0,31	0,781
Siamese+PTPARANMT	300	300	2	**0,24**	**0,828**
para-multi-MiniLM-L12-v2	768	3072	12	2,27	0,761
dpr-ctx-enc-bert-base-multi	768	3072	12	0,56	0,560

Tabela 8: Modelos de similaridade sentencial e resultados obtidos.

Modelo	Submissão	MSE	Pearson
Siamese+PTPARANMT	-	**0,24**	**0,828**
Stilingue (Fonseca & Alvarenga, 2019)	3	0,47	0,817
L2F	BL	0,52	0,778
IPR (Rodrigues et al., 2019b)	1	0,52	0,826
ASAPPy (Santos et al., 2019)	1-ptbr	0,58	0,730
DeepLearning Brasil (Rodrigues et al., 2019a)	Ensemble	0,59	0,785
NILC (Cabezudo et al., 2020)	2	0,64	0,729
baseline	Overlap	0,75	0,577
PUCPR (de Souza et al., 2019)	comNILC	0,85	0,678
LIACC	2	1,02	0,459
baseline	BoW-2	1,15	0,175

Tabela 9: Comparação do modelo Siamese+PTPARANMT com participantes da iniciativa *ASSIN* 2.

Modelo	Tamanho	Neurônios	Pré-treinamento?
S2S	200	200	Não
paraPTT5	768	3072	Sim
REF	–	–	Não
Cópia	–	–	Não
Ingênuo	–	–	Não

Tabela 10: Configurações dos modelos neurais do tipo *sequência-para-sequência*.

Para selecionar a melhor configuração de parâmetros de filtro do córpus, foi utilizado um modelo neural, aqui denominado *PTT5finne*, com pesos ajustados a partir do modelo público *PTT5-base* (Carmo et al., 2020). Esse modelo passou por iterações com os parâmetros de sobreposição de texto e *para-score* em um esquema de busca em *grid*, visando maximizar a métrica de similaridade semântica *bert-score*. A escolha das métricas *para-score* e sobreposição como filtros foi fundamentada nos resultados apresentados em (Krishna et al., 2020). No referido estudo, propôs-se o treinamento de um modelo de reescrita de paráfrases em inglês (STRAP) a partir de um subconjunto filtrado de pares artificiais de paráfrases, utilizando essas métricas como critério. Além disso, a adição da métrica *bert-score* foi motivada por sua significativa correlação com a avaliação humana de textos em nível sentencial (Zhang et al., 2020). A Figura 1 ilustra o mapa de calor do *bert-score* obtido por meio desse processo de busca (melhor visualizado em formato eletrônico).

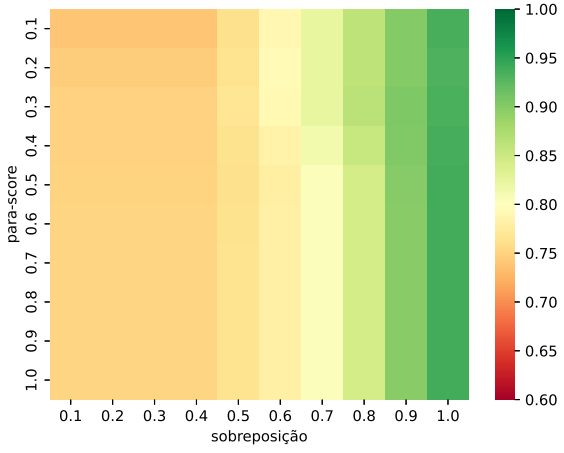

Figura 1: *bert-score* para diferentes versões do córpus *PTPARANTM* com base nas medidas de para-score e sobreposição.

O mapa de calor na Figura 1 revela que o modelo *PTT5finne* alcança 0,775 pontos de *bert-score* no intervalo de valores (0,6:0,6). Em outras palavras, nos respectivos intervalos das métricas de sobreposição e *para-score*, o modelo *PTT5finne* apresentou o melhor desempenho em termos de correlação semântica (*bert-score*) entre o texto gerado e o texto do conjunto-alvo. Estes intervalos são similares aos apresentados para as métricas de *para-score* e sobreposição em Krishna et al. (2020). A filtragem dos dados do córpus com base neste parâmetro resultou em 73.476 pares de paráfrase. Este conjunto foi particionado em uma porção de desenvolvimento de 90% (59.584 pares de sentenças), e 10% de teste (7.347 pares). Do conjunto de desenvolvimento, uma porção de 5% (3.306 pares) foi utilizada para validação e iteração dos hiper-parâmetros, e o restante foi utilizado para treino.

5.2. Avaliação

Os modelos propostos — *S2S* e *paraPTT5* — foram treinados por *back-propagation*, e sua capacidade de generalização foi avaliada em relação ao conjunto de testes. O modelo*S2S* foi otimizado pelo método *adagrad* com uma taxa inicial de 0,15 por um total de 200 mil épocas, e o modelo *paraPTT5* foi otimizado pelo método *adam* com uma taxa inicial de 0,00005 por um total de 2 épocas. Em ambos os casos, seguimos a configuração adotada em Kaplan et al. (2020) definindo um tamanho de *batch* com 20 pares de frases, e limitando as sequências de entrada e saída a sentenças de tamanho 50. Por fim, os dois modelos foram otimizados com o objetivo de minimizar o erro médio das sub-palavras geradas em relação às sub-palavras alvo, ou entropia cruzada.

Os modelos *S2S* e *paraPTT5* foram submetidos à avaliação intrínseca e humana. Além disso, no caso da avaliação intrínseca, estes modelos foram comparados também aos sistemas de *baseline REF*, *Cópia* e *Ingênuo*. As duas modalidades de avaliação são discutidas individualmente nas próximas seções.

5.2.1. Avaliação intrínseca

A avaliação intrínseca realizada foi baseada em cinco métricas de qualidade dos textos gerados pelos modelos em relação aos texto-alvo do conjunto de teste. Três destas métricas são de motivação computacional, a saber: *bert-score*, *ROUGE* e *para-score*. Para *ROUGE* e *bert-score*, utilizamos as implementações disponíveis na biblioteca *Transformers Hugging Face* (Wolf et al., 2020). Os pesos pré-treinados da métrica *bert-score* foram obtidos a partir da versão pública *bert-base-multilingual-cased* do modelo.

A estas métricas, foram acrescidas duas medidas de complexidade da escrita discutidas em Leal et al. (2023): a métrica de complexidade lexical Brunet e a métrica de complexidade sintática da distância do grafo. A escolha dessas duas métricas foi motivada pelos resultados apresentados em Leal et al. (2018) para a tarefa de identificação da complexidade textual, e pela facilidade de replicação das mesmas para o português. Em ambos os casos, optamos por reimplementar o código apresentado em Leal et al. (2023) com uso do pacote *spaCy* em Python.

A métrica Brunet relaciona a taxa dos erros tipográficos com o tamanho do texto, e apresenta valores típicos entre 10 e 20, sendo que um texto mais rico (e complexo) produz valores menores. A distância do grafo estima a complexidade sintática de uma sentença representada na forma de um grafo de dependências considerando a relação entre suas palavras e a distância do arco de dependência entre elas, apresentando valores maiores (indicativos de maior complexidade) à medida que as distâncias de dependência aumentam.

A Tabela 11 apresenta os resultados obtidos para a tarefa de reescrita sentencial com base nas métricas selecionadas. O melhor resultado obtido por um dos dois modelos propostos — seja *S2S* ou *paraPTT5* — de acordo com cada métrica de avaliação é destacado em negrito.

Os resultados da Tabela 11 sugerem que, de modo geral, os modelos propostos *S2S* e *paraPTT5* são superiores às alternativas. O modelo *paraPTT5* é superior para as métricas de complexidade sentencial (complexidade de distância do grafo e leiturabilidade Brunet), enquanto que o modelo *S2S* se destacou nas métricas *para-score*, *bert-score* e *ROUGE*. Além disso, observa-se que o *baseline REF*, que produz sempre a saída esperada (e assim obtém valores máximos para as métricas de similaridade textual como *para-score*, *bert-score*, etc.) apresenta resultados inferiores aos dos modelos propostos para todas as demais métricas por falta de diversidade linguística, ilustrando a necessidade de equilíbrio entre múltiplos critérios conflitantes para o sucesso da tarefa de reescrita.

5.2.2. Avaliação humana

Em complemento à avaliação intrínseca discutida na seção anterior, foi realizada também uma análise humana dos textos gerados pelos mode-

	Modelos				
Métrica	S2S	paraPTT5	REF	Cópia	Ingênuo
para-score	**88,7**	88,1	100,0	81,4	29,9
bert-score	**91,7**	90,6	100,0	87,6	68,3
ROUGE-1	**76,5**	70,2	100,0	54,8	23,3
ROUGE-2	**61,1**	54,4	100,0	30,8	5,8
distância do grafo	27,9	**22,9**	29,1	29,7	48,6
Brunet	5,5	**5,2**	5,5	5,5	6,5

Tabela 11: Avaliação intrínseca dos textos gerados.

los *S2S* e *paraPTT5*. Para este fim, oito falantes nativos da língua portuguesa foram solicitados a avaliar um total de 102 pares de paráfrases provenientes de uma amostragem do conjunto de teste do córpus quanto ao grau de relação de paráfrase do par, e também com relação ao grau de fluência de cada texto individualmente. Ambas avaliações seguiram as mesmas diretrizes discutidas na seção 3.2, ou seja, atribuindo-se escores de 0 a 2 a cada par de frases representado o grau da relação de paráfrase entre elas, e escores 0 a 2 a cada frase individual representando seu nível de fluência.

Foi alcançada uma concordância entre anotadores de $0,88$ de acordo com o índice kappa de Cohen (Landis & Koch, 1977). Os pares de frases foram agrupados em intervalos de 10 com base na métrica de similaridade de paráfrases *para-score*, e os resultados na Tabela 12 sumarizam o número de vezes que cada escore de paráfrase e fluência (Fluência.O e Fluência.A de textos Origem e Alvo, respectivamente) foi escolhido pelos avaliadores das 102 amostras.

Conforme observado na Tabela 12, os pares ruidosos de ambos os modelos estão majoritariamente confinados aos primeiros dois intervalos $((0.23, 0.672), (0.672, 0.766))$, o que corresponde a 30 - 46% de todos os pares que possuem uma forte relação de paráfrase. Além disso, no intervalo superior, 86,63% (para *S2S*) e 90% (para *paraPTT5*) dos pares possuem uma forte relação de paráfrase.

Em complemento a estes resultados, a Tabela 13 apresenta os valores médios para as métricas de fluência e grau de relação de paráfrase de cada par obtidos pelos dois modelos avaliados.

A Tabela 13 indica que o modelo *paraPTT5* obtém os melhores resultados tanto com base na fluência textual quanto na qualidade das paráfrases geradas, o que pode ser tomado como indício de que LLMs são efetivamente úteis para a tarefa de transferência de estilo textual por meio de paráfrases, e coloca assim a questão de qual seria o desempenho do modelo *paraPTT5* sem o refinamento dos pesos a partir do LLM. Em uma análise *post hoc* (aqui não detalhada) na qual o modelo *paraPTT5* foi substituído pelo modelo *S2S*, observamos que *paraPTT5* seria a melhor alternativa, confirmando assim a superioridade da técnica de transferência de aprendizado com LLMs.

5.2.3. Exemplos de textos gerados

Como forma de exemplificar a qualidade dos pares de paráfrases gerados pelo modelo *paraPTT5*, a Tabela 14 apresenta amostras aleatórias de textos-origem e suas paráfrases. Para facilitar a visualização, os exemplos são agrupados informalmente em três categorias do grau de erro (baixa, média e alta) conforme o grau de erro na relação de paráfrase de cada par avaliado.

Finalmente, a Tabela 15 apresenta amostra de textos-origem gerados para exemplificar a fluência dos textos gerados através do modelo paraPTT5. Os exemplos são divididos informalmente em três categorias do grau de erro (baixa, média e alta), de acordo com a pontuação de fluência do texto.

Os exemplos apresentados nas Tabelas 14 e 15 apresentam uma série de erros de natureza semântica e superficial proporcionais ao grau de erro reportado. Estes erros são, em grande parte, decorrentes do método de criação e filtragem do córpus *PTPARANMT* (seções 3 e 5.1). Em especial, observamos que o método de tradução reversa, por se tratar de um processo automático, pode ter introduzido de forma não intencional ruídos durante o processo de produção do estilo-alvo desejado, como já relatado em trabalhos similares para a língua inglesa Wieting & Gimpel (2018); Krishna et al. (2020). Além disso, observamos que no processo de filtragem dos dados foram priorizados os pares de sentenças de maior *para-score*, o que privilegia a preservação da semântica com certo detrimento à forma superficial.

Modelo	Intervalo para-score	# Pares	Fluência.O 0	1	2	Fluência.A 0	1	2	Paráfrase 0	1	2
S2S	(0.23, 0.672)	10	13	5	2	12	4	4	6	11	3
	(0.672, 0.766)	10	9	7	4	3	8	9	6	6	8
	(0.766, 0.827)	10	9	7	4	4	5	11	8	11	1
	(0.827, 0.926)	10	14	5	1	8	8	4	11	9	0
	(0.926, 1.0)	10	9	5	6	9	6	5	18	1	1
paraPTT5	(0.207, 0.638)	11	11	4	7	13	3	6	8	10	4
	(0.638, 0.747)	10	12	7	1	10	7	3	9	7	4
	(0.747, 0.828)	10	13	5	2	10	7	3	12	8	0
	(0.828, 0.896)	10	11	5	4	10	5	5	10	8	2
	(0.896, 1.0)	11	12	8	2	12	8	2	19	3	0

Tabela 12: Avaliação humana dos textos gerados.

Modelo	Fluência	Paráfrase
S2S	75,0%	87,0%
paraPTT5	83,2%	90,4%

Tabela 13: Resultados médios de fluência e paráfrase.

5.3. Considerações

Os resultados da avaliação dos modelos *S2S* e *paraPTT5* permitem uma série de observações. Em primeiro lugar, destaca-se a importância do ajuste de pesos baseado em LLMs e da seleção criteriosa de paráfrases para compor o conjunto de dados, sem os quais o texto gerado não seria minimamente aceitável do ponto de vista de um leitor humano.

Em segundo lugar, observa-se que a técnica de transferência de conhecimento permitiu ao modelo *paraPTT5* produzir textos mais coesos e semelhantes a paráfrases se comparado ao modelo *S2S*. Uma possível explicação para esse resultado é que *S2S* é um modelo neural sequencial do tipo *transformer* que não utiliza transferência de conhecimento. Assim, a correlação positiva observada nos resultados do modelo *paraPTT5* no conjunto de dados de teste parece estar relacionada à qualidade das representações internas dos LLMs.

6. Conclusões

Este trabalho apresentou um primeiro estudo em transferência de estilo textual arbitrário utilizando paráfrases em Português, tratando da construção do córpus de paráfrases *PTPARANMT* e *embeddings* de mesmo nome, e do uso destes recursos na tarefa de reescrita sentencial baseada em paráfrases.

Os recursos linguístico-computacionais construídos[3] foram inicialmente validados nas tarefas de simplificação e similaridade sentencial, obtendo resultados superiores aos das alternativas consideradas. No caso da tarefa de similaridade sentencial, os resultados obtidos foram inclusive superiores aos reportados na área tomando-se por base o *benchmark ASSIN* 2 (Real et al., 2020).

Uma vez estabelecidos estes resultados iniciais, o córpus *PTPARANMT* foi então empregado na tarefa para a qual havia sido realmente desenvolvido, ou seja, a transferência de estilo arbitrário em Português. Foram propostos para esse fim dois modelos principais, sendo um baseado na arquitetura *sequência-para-sequência* e o outro obtido pelo refino de um grande modelo de língua existente. Os modelos propostos foram avaliados de forma intrínseca e com auxílio de juízes humanos, sugerindo-se a importância do ajuste de pesos baseado em LLMs e da seleção criteriosa de paráfrases para compor o conjunto de treino, e a superioridade da técnica de transferência de aprendizado a partir de LLMs em relação à arquitetura *sequência-para-sequência* na tarefa em questão.

O presente estudo deixa diversas oportunidades de trabalho futuro. Por exemplo, uma alternativa relevante à presente abordagem, e que tem se popularizado rapidamente nas áreas do PLN e GLN, seria o uso de técnicas de engenharia de *prompts* e aprendizado *few-shoot* com uso de LLMs (Min et al., 2023). Estudos recentes de transferência de estilo, como em Troiano et al. (2023), têm demonstrado ganho significativo em relação ao uso de modelos pré-treinados em cenários onde o estilo-alvo é arbitrário ou desconhecido (Krishna et al., 2022; Reif et al., 2022). No entanto, por ser uma inovação científica re-

[3] https://github.com/pablocosta/paperLinguamaticaTSTBR

Erro	Texto original	Paráfrase
baixo	*aqui está nossa revisão do estilo de vida de wall street*	*aqui está nossa avaliação sobre walls-treet lifestyle*
médio	*assim no modelo de privatização pura estado interferiria menos no seb exceto no que considerava privatização*	*assim no modelo puro de privatização estado interferia menos no seb com exceção do que*
alto	*ao adicionar 75 por cento em cima dos seus gastos e comprar árvores por essa quantidade tudo que você gasta é devol-vido você*	*ao adicionar 75 ao exceder os gastos e comprar árvores por essa quantia todos os gastos são*

Tabela 14: Amostras aleatórias de textos produzidos pelo modelo paraPTT5 com diferentes graus de erro de paráfrase.

Erro de fluência	Texto gerado
Baixo	*acesse fórum de viagens do tripadvi-sor sobre portland e faça perguntas ao*
Médio	*alimentação no mercado central mercado central 27 km*
Alto	*inscreverse para atualizações seja notificado quando atualizarmos in-formações sobre vgoc*

Tabela 15: Amostras aleatórias de textos gerados pelo modelo paraPTT5 com diferentes graus de erro de fluência.

cente, métodos deste tipo ainda apresentam desempenho inferior ao das abordagens que fazem refinamento de LLMs pré-treinados nos casos em que um conjunto de dados de proporção significativa esteja disponível (Scao & Rush, 2021; Puri et al., 2023; Liu et al., 2023), havendo portanto oportunidade para mais pesquisas.

Além disso, consideramos também a construção de um córpus de paráfrases composto de estilos reais (i.e., produzidos por diferentes autores humanos) no lugar de um conjunto de dados sintético como o presente córpus *PTPARANMT*. Uma iniciativa deste tipo proporcionaria não apenas um maior grau de realismo à tarefa computacional, mas poderia também auxiliar na redução do ruído proveniente do método de tradução automática aqui empregado.

Agradecimentos

O segundo autor conta com apoio do processo nro. 2021/08213-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

Agirre, Eneko, Daniel Cer, Mona Diab & Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. Em *1st Joint Conference on Lexical and Computational Semantics*, 385–393.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. ArXiv [cs.CL]. 10.48550/arXiv.1409.0473.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins & Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. Em *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4555–4567. 10.18653/v1/2020.acl-main.417.

Beddiar, Djamila Romaissa, Md Saroar Jahan & Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media* 24. 100153. 10.1016/j.osnem.2021.100153.

Bhagat, Rahul & Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics* 39(3). 463–472. 10.1162/COLI_a_00166.

Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press. 10.1017/CBO9780511814358.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. Em *34th International Conference on Neural Information Processing Systems*, 1877–1901.

Cabezudo, Marco Antonio Sobrevilla, Marcio Lima Inácio, Ana Carolina Rodrigues, Edresson Casanova & Rogério Figueredo de Souza. 2020. Nilc at assin 2: exploring multilingual approaches. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 49–58.

Carlson, Keith, Allen Riddell & Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science* 5(10). 171920. 10.1098/rsos.171920.

Carmo, Diedre, Marcos Piau, Israel Campiotti, Rodrigo Nogueira & Roberto Lotufo. 2020. PTT5: Pretraining and validating the PT5 model on Brazilian Portuguese data. ArXiv [cs.CL]. 10.48550/arXiv.2008.09144.

Dong, Chenhe, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen & Min Yang. 2022. A survey of natural language generation. *ACM Computing Surveys* 55(8). 173. 10.1145/3554727.

Fonseca, Evandro & João Paulo Reis Alvarenga. 2019. Wide and deep transformers applied to semantic relatedness and textual entailment. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 68–77.

Gatt, Albert & Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61(1). 65–170.

Glorot, Xavier & Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. Em *13th International Conference on Artificial Intelligence and Statistics (AIStat)*, 249–256.

Goldberg, Yoav. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57. 345–420.

Gonçalo Oliveira, Hugo & Ana Alves. 2021. AIA-BDE: um corpo de perguntas, variações e outras anotações. *Linguamática* 13(2). 19–35. 10.21814/lm.13.2.350.

Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2016. *Deep learning*. MIT Press.

Hartmann, Nathan S., Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues & Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 122–131.

Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari & Trevor Cohn. 2018. Iterative back-translation for neural machine translation. Em *2nd Workshop on Neural Machine Translation and Generation*, 18–24. 10.18653/v1/W18-2703.

Hu, Zhiqiang, Roy Ka-Wei Lee, Charu C. Aggarwal & Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter* 24(1). 14–45. 10.1145/3544903.3544906.

Hu, Zhiting, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov & Eric P Xing. 2017. Toward controlled generation of text. Em *International Conference on Machine Learning*, 1587–1596.

Jin, Di, Zhijing Jin, Zhiting Hu, Olga Vechtomova & Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics* 48(1). 155–205. 10.1162/coli_a_00426.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu & Dario Amodei. 2020. Scaling laws for neural language models. ArXiv [cs.LG]. 10.48550/arXiv.2001.08361.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen & Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. 10.18653/v1/2020.emnlp-main.550.

Kingma, Diederik P. & Jimmy Ba. 2015. Adam: A method for stochastic optimization. Em *3rd International Conference on Learning Representations (ICLR)*, 10.48550/arXiv.1412.6980.

Krishna, Kalpesh, Deepak Nathani, Xavier Garcia, Bidisha Samanta & Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. Em *60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7439–7468. 10.18653/v1/2022.acl-long.514.

Krishna, Kalpesh, John Wieting & Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. Em *Empirical Methods in Natural Language Processing (EMNLP)*, 737–762. 10.18653/v1/2020.emnlp-main.55.

Kudo, Taku & John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 66–71. 10.18653/v1/D18-2012.

Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33. 159–174. 10.2307/2529310.

Leal, Sidney Evaldo, Magali Sanches Duran & Sandra Maria Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. Em *27th International Conference on Computational Linguistics (COLING)*, 401–413.

Leal, Sidney Evaldo, Magali Sanchez Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2023. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation* 10.1007/s10579-023-09693-w.

Li, Juncen, Robin Jia, He He & Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. ArXiv [cs.CL]. 10.48550/arXiv.1804.06437.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi & Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55(9). 10.1145/3560815.

Luo, Fuli, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui & Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. ArXiv [cs.CL]. 10.48550/arXiv.1905.10060.

Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz & Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys* 56(2). 1–40. 10.1145/3605943.

Naseem, Usman, Imran Razzak, Shah Khalid Khan & Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing* 20(5). 1–35. 10.1145/3434237.

Pennington, J., R. Socher & C. D. 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.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. Em *3rd Conference on Machine Translation*, 186–191. 10.18653/v1/W18-6319.

Puri, Ravsehaj Singh, Swaroop Mishra, Mihir Parmar & Chitta Baral. 2023. How many data samples is an additional instruction worth? Em *Findings of the Association for Computational Linguistics: EACL*, 1042–1057. 10.18653/v1/2023.findings-eacl.77.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li & Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21(140). 1–67.

Rao, Sudha & Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 129–140. 10.18653/v1/N18-1012.

Real, Livy, Erick Fonseca & Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: a quick overview. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 406–412. 10.1007/978-3-030-41505-1_39.

Reif, Emily, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch & Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. Em *60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 837–848. 10.18653/v1/2022.acl-short.94.

Reimers, Nils & Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. Em *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. 10.18653/v1/D19-1410.

Riley, Parker, Noah Constant, Mandy Guo, Girish Kumar, David Uthus & Zarana Parekh. 2021. TextSETTR: Few-shot text style extraction and tunable targeted restyling. Em *59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3786–3800. 10.18653/v1/2021.acl-long.293.

Rodrigues, Ruan Chaves, Jéssica Rodrigues da Silva, Pedro Vitor Quinta de Castro, Nádia Silva & Anderson da Silva Soares. 2019a. Multilingual transformer ensembles for Portuguese natural language tasks. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 27–38.

Rodrigues, Rui, Paula Couto & Irene Rodrigues. 2019b. IPR: The semantic textual similarity and recognizing textual entailment systems. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 39–48.

Santos, José, Ana Alves & Hugo Gonçalo Oliveira. 2019. ASAPPpy: a Python framework for Portuguese STS. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 14–26.

Scao, Teven Le & Alexander M. Rush. 2021. How many data points is a prompt worth? Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2627–2636. 10.18653/v1/2021.naacl-main.208.

Shen, Tianxiao, Tao Lei, Regina Barzilay & Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems* 30.

Shliazhko, Oleh, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova & Tatiana Shavrina. 2022. mGPT: Few-shot learners go multilingual. ArXiv [cs.CL]. 10.48550/arXiv.2204.07580.

de Souza, João Vitor Andrioli, Lucas E. S. Oliveira, Yohan Bonescki Gumiel, Deborah Ribeiro de Carvalho & Cláudia Maria Cabral Moro. 2019. Incorporating multiple feature groups to a siamese neural network for semantic textual similarity task in Portuguese texts. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, 59–68.

Suzgun, Mirac, Luke Melas-Kyriazi & Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2195–2222. 10.18653/v1/2022.emnlp-main.141.

Tiedemann, Jörg & Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. Em *22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, 479–480.

Troiano, Enrica, Aswathy Velutharambath & Roman Klinger. 2023. From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering* 29(4). 849–908. 10.1017/S1351324922000407.

Wang, Yunli, Yu Wu, Lili Mou, Zhoujun Li & Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. Em *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3573–3578. 10.18653/v1/D19-1365.

Warstadt, Alex, Amanpreet Singh & Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7. 625–641. 10.1162/tacl_a_00290.

Wieting, John & Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. Em *56th Annual Meeting of the Association for Computational Linguistics*, 451–462. 10.18653/v1/P18-1042.

Wieting, John, Kevin Gimpel, Graham Neubig & Taylor Berg-Kirkpatrick. 2021. Paraphrastic representations at scale. ArXiv [cs.CL]. 10.48550/arXiv.2104.15114.

Wieting, John, Jonathan Mallinson & Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. Em

Conference on Empirical Methods in Natural Language Processing (EMNLP), 274–285. doi 10.18653/v1/D17-1026.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. doi 10.18653/v1/2020.emnlp-demos.6.

Xu, Wei, Alan Ritter, Bill Dolan, Ralph Grishman & Colin Cherry. 2012. Paraphrasing for style. Em *International Conference on Computational Linguistics (COLING)*, 2899–2914.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger & Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. Em *International Conference on Learning Representations*, on–line.

Zhang, Xiang, Junbo Zhao & Yann LeCun. 2015. Character-level convolutional networks for text classification. Em *28th International Conference on Neural Information Processing Systems*, 649–657.

Propuesta recibida en julio 2023 y aceptada para publicación en noviembre 2023.

Detección de operadores modales: una primera exploración en castellano

Detection of modal operators: first explorations in Spanish

Javier Obreque
Pontificia Universidad Católica de Valparaíso

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso

Resumen

El artículo presenta una propuesta metodológica de carácter mixto — con énfasis en el aspecto cuantitativo — para la detección y registro de operadores modales. Estas unidades pueden definirse como un amplio y heterogéneo conjunto de expresiones que se utilizan en la comunicación escrita y oral para imprimir la visión subjetiva del emisor en su propio enunciado. La presente propuesta se basa en la explotación de un corpus paralelo para aumentar con medios cuantitativos un listado inicial de ejemplos obtenido en una etapa cualitativa. La metodología es simple, efectiva e independiente de lengua, aunque en este primer ensayo nos enfocamos en el castellano.

Palabras clave

modalización, operadores modales, métodos mixtos, corpus paralelo

Abstract

This article presents a mixed methods approach — with emphasis on the quantitative side — for the detection and recording of modal operators. These units are defined as a broad and heterogeneous set of expressions used in written and oral communication to imprint the subjective vision of the writers/speakers in their own utterance. The present proposal is based on the exploitation of a parallel corpus to augment with quantitative means an initial list of examples obtained in a qualitative stage. The methodology is simple, effective, and language independent, although in this first test we focus on the Spanish language.

Keywords

modality, modal operators, mixed methods, parallel corpora

1. Introducción

El estudio de la modalización tiene una larga tradición en las ciencias del lenguaje, que comienza con la lógica de Aristóteles y alcanza su maduración en la filosofía medieval. Los lógicos escolásticos ya distinguían, en una proposición, entre el *dictum* y el *modus* (Ridruejo, 1999). El primer término se refiere al contenido proposicional del enunciado y el segundo especifica un tipo de modalización.

(1) *El gato está sobre la mesa.*

(2) *Parece que el gato está sobre la mesa.*

En el ejemplo (1) se presenta una oración declarativa, descriptiva, sobre un estado de cosas. En (2), en cambio, se expone un enunciado más complejo, en que ese mismo contenido proposicional (*dictum*) aparece subordinado a una marca de modalización (*modus*), y el efecto inmediato es que se comunica el menor grado de certeza del hablante.

El interés de la lingüística por la modalización comenzó a tomar impulso principalmente a partir de la teoría de la enunciación de tradición francesa (Benveniste, 1966, 1974; Kerbrat-Orecchioni, 1987). A partir de estas investigaciones se ha definido la modalización como parte de los rastros de la subjetividad del sujeto productor del discurso y de la situación en la que tiene lugar el acto enunciativo original (Charaudeau, 1994). Se conserva la distinción medieval entre el *dictum*, que presenta hechos o datos con objetividad, y el *modus*, que señala la actitud, posicionamiento, certeza, subjetividad, opiniones, puntos de vista, sentimientos o emociones del emisor, ya sea ante el oyente o ante el contenido de su propio enunciado (Otaola, 1988; Palmer, 2001; Wiebe et al., 2005; Miwa et al., 2012; Taboada, 2016).

En el presente artículo denominamos *operadores modales* (OM) a estas marcas que manifiestan la subjetividad del hablante sobre su enunciado.

DOI: 10.21814/lm.15.2.411

Los OM son un fenómeno universal, ya que al parecer se encuentran en todas las lenguas (Nissim et al., 2013). El problema es que, desde el punto de vista lingüístico, son difíciles de delimitar conceptualmente. No pertenecen a una sola categoría gramatical, ni a una clase cerrada de palabras, ni a unidades puramente funcionales (Blanché, 1966; Lozano et al., 1982; Nuyts, 2005; Müller, 2007; Nuyts, 2016b,a). Se trata, más bien, de una clase abierta que designa un dominio de tipo conceptual y con medios de expresión muy variados, incluyendo elementos poliléxicos (Pottier, 1977; Calsamiglia & Tusón, 1999; Narrog, 2012). Pueden incluirse en este conjunto expresiones tan diversas como *evidentemente, de hecho, debe tener, es posible que, afortunadamente*, etc. Esta heterogeneidad dificulta, por supuesto, la creación y mantenimiento de inventarios exhaustivos (Ruppenhofer & Rehbein, 2012; Marasović et al., 2016; Pyatkin et al., 2021).

El objetivo del presente trabajo es desarrollar un método computacional que permita identificar, extraer y clasificar de manera automática un amplio conjunto de OM de una lengua. Si bien, como decíamos, el inventario completo es muy difícil de obtener, sí tiene sentido al menos intentar el registro de las unidades más frecuentemente utilizadas.

El método que proponemos para ello está basado en el análisis estadístico de un corpus paralelo, lo que implica una mirada interlingüística en que se utiliza como medio de análisis una lengua distinta a la analizada. Otra característica es que, si bien es esencialmente cuantitativo, también puede considerarse de enfoque mixto, ya que incluye una fase cualitativa previa en la que hay un proceso de anotación manual de corpus.

Los resultados preliminares que se muestran en este artículo corresponden a la lengua castellana. Sin embargo, el valor de la investigación está en que, al tratarse de una metodología principalmente cuantitativa, puede ser aplicada a cualquier lengua que disponga de corpus paralelos y de un conjunto inicial de ejemplos de OM, que puede incrementarse gradualmente con la aplicación del mismo método.

El artículo se organiza de la siguiente manera: después de esta introducción, en la sección 2 se presenta un breve marco teórico que explica la definición de los OM como objeto de estudio (2.1), sus categorías principales (2.2), así como la relación entre el registro de OM y la anotación manual de modalizadores (2.3). En la sección 3 se presenta la propuesta metodológica, los resultados en la sección 4 y en la última sección (5) se presentan las conclusiones y el trabajo todavía por realizar. Además, un sitio web[1] acompaña a este artículo, ofreciendo documentación detallada, el código fuente utilizado y los datos de entrada y salida del proceso.

2. Marco teórico

2.1. Operadores modales

Hemos definido los OM como operadores de clase abierta que marcan un tipo de modalización en un enunciado. A continuación, complementamos esta caracterización con una descripción de sus propiedades esenciales.

Una de las propiedades esenciales de los OM es el aumento del poder expresivo y, a la vez, de la complejidad sintáctica y semántica de los enunciados en los que operan (Van Dijk, 1980, 2012; Fuentes Rodríguez, 2003). En la gramática y la lingüística textuales (Bernárdez, 1982; Casado Velarde, 1993; De Beaugrande & Dressler, 1997; Cuenca, 2010), se denomina operador una unidad que actúa de esta manera, es decir, que es exterior a la estructura sintáctica de la cláusula a la que afecta (Fuentes Rodríguez, 2009). Con frecuencia el *modus* es un elemento parentético o, más típicamente, el *dictum* se encuentra en una estructura subordinada.

Algunos autores han identificado a los OM como un tipo de marcador discursivo (Martín-Zorraquino & Portolés, 1999), pero existen algunos argumentos para rechazar esta categorización. Si bien en algunos contextos pueden cumplir ambos roles, un OM en principio no es un marcador discursivo porque su función no es orientar intratextualmente la generación de inferencias que un potencial lector/interlocutor debe hacer durante la comprensión de una unidad comunicativa. La función que cumplen los OM es, en cambio, el resultado de una relación extratextual o exofórica, es decir, la que indica cuál es la posición que toma el sujeto productor del mensaje ante su propio enunciado (Barrenechea, 1979).

En particular, los OM se diferencian de los conectores porque estos, aunque también exceden los límites sintácticos de la oración, actúan como enlace entre dos enunciados diferentes. El OM, en cambio, no conecta sino que, como dijimos, opera en un mismo enunciado (Fuentes Rodríguez, 2009). Nuevamente es necesario aclarar que, en ciertas circunstancias, un OM también puede cumplir funciones de un conector, y viceversa. Este es el caso, por ejemplo, de la expresión *de hecho*, que puede funcionar como un OM, pero también, en algunos contextos, como un co-

[1] http://www.tecling.com/moper

nector justificativo (Fuentes Rodríguez, 2009) o bien como un operador de refuerzo argumentativo (Martín-Zorraquino & Portolés, 1999). Este sería, sin embargo, solamente un caso de polifuncionalidad (der Auwera & Ammann, 2005).

Los OM han sido objeto de interés de distintas corrientes de la lingüística del texto porque son útiles para la caracterización de los textos. Por ejemplo, algunos investigadores se interesan en medir el tipo de modalización de los textos académicos (Gutiérrez, 2010), describir modalizadores específicos de un determinado género (Sologuren & Venegas, 2022) o describir expresiones modales como marcadores metadiscursivos (Salas, 2015).

2.2. Tipología de operadores modales

Las tres categorías fundamentales de OM proceden de la lógica clásica: epistémica, deóntica y alética. Los OM epistémicos indican el grado de compromiso y conocimiento en los enunciados (Pérez Canales, 2009; González et al., 2016). Tradicionalmente, se han vinculado con los verbos *saber* y *creer* (*pensamos, intuimos, conocemos*), pero también con otras categorías gramaticales, como los adverbios (*aparentemente, supuestamente*, etc.). Los OM deónticos, por su parte, manifiestan la expresión de una obligación (Van Dijk, 1980; Nuyts, 2005) (*se debe, es necesario que*). En cuanto a los OM aléticos, estos expresan necesidad o posibilidad (Van Dijk, 1980) (*es posible que, probablemente*). Esta es una categoría menos utilizada en análisis lingüísticos (Nuyts, 2006), pero está en la base de las dos anteriores (Calsamiglia & Tusón, 1999).

Esta tipología inicial no agota las posibilidades de estudio de las expresiones modales (Narrog, 2012), pero constituye al menos una base útil (Müller, 2007; Portner, 2009). Existen, además, otras categorías que no están del todo asentadas en el análisis del discurso: veredictorias o veredictivas (acerca de la verdad o la mentira de los enunciados), valorativas o axiológicas (evaluación positiva o negativa), volitivas o bulomayeicas (deseo, preferencia o necesidad), de usualidad, de cantidad (Greimas, 1973; Lozano et al., 1982; Otaola, 1988; der Auwera & Plungian, 1998; Calsamiglia & Tusón, 1999; Nuyts, 2006; Cuenca, 2010). A estas pueden sumarse otras incluso menos estables que algunos autores disponen bajo la categoría de indeterminadas (Kalinowski, 1976; Lozano et al., 1982; Portner, 2009).

2.3. Registro de operadores modales: anotación manual y automática

Aunque, como ya se anticipó, la heterogeneidad de los OM hace imposible su consideración como clase cerrada de palabras, hay estudios que se han propuesto identificarlos e inventariarlos (Pyatkin et al., 2021). Existen también recopilaciones enfocadas específicamente en verbos modales (Brandt, 1999; Ruppenhofer & Rehbein, 2012; Marasović et al., 2016, entre otros), es decir, formas verbales que prototípicamente expresan las nociones propias de las modalizaciones del enunciado, como *sé* o *creemos* (modalidad epistémica), *debes* o *haz* (deóntica), etc.

En lengua castellana también se han llevado a cabo esfuerzos destacables por desarrollar inventarios o registros de OM, y tres de ellos corresponden a proyectos lexicográficos de partículas discursivas, como el *Diccionario de partículas* (Santos Río, 2003), el *Diccionario de partículas discursivas del español*[2] y el *Diccionario de conectores y operadores del español* (Fuentes Rodríguez, 2009).

El análisis de los OM también se ha llevado a cabo a través de estudios de corpus (Hendrickx et al., 2012; Nissim et al., 2013; Ghia et al., 2016). En este tipo de estudios la metodología se basa en la anotación manual de corpus, que sirve para incrementar los registros de OM y como recurso para desarrollar propuestas computacionales de detección y anotación de estas partículas (Baker et al., 2010; Rubinstein et al., 2013; Quaresma et al., 2014). Estos registros pueden ser utilizados de forma directa o bien como fuente para crear sistemas de detección basados en reglas (Saurí et al., 2005, 2006; Soni et al., 2014; Lee et al., 2015).

3. Metodología

La presente propuesta metodológica se sustenta en la estrategia de uso de una segunda lengua para el estudio de un aspecto particular de una lengua determinada. Para este primer ensayo, la lengua objetivo es el castellano, y la segunda lengua el inglés, por cuestiones prácticas de disponibilidad de material. En lo esencial, la metodología que describimos en este artículo es la de un proceso de clasificación, en el que a partir de un listado de expresiones (todas las palabras y secuencias de palabras de un corpus), pretendemos clasificarlas en las categorías de OM y no-OM. Según nuestro

[2]Diccionario de Partículas Discursivas del Español, Briz, A. and Pons, S. and Portolés, J. (coords.), `http://www.dpde.es`.

diseño de investigación, se requiere un listado de OM en la segunda lengua, y se requiere aplicar la metodología en las dos direcciones, es decir, invirtiendo la segunda vez el orden de lengua objetivo/medio (castellano/inglés). Además, el proceso debe realizarse de manera independiente por cada categoría de OM. En esta ocasión hemos probado con OM epistémicos, deónticos, aléticos y valorativos. A continuación presentamos los materiales utilizados para esta aplicación metodológica (3.1) y, seguidamente, las fases o etapas del procedimiento (3.2).

3.1. Materiales

El insumo principal de nuestra metodología es un corpus paralelo (CP). Los CP han sido utilizados ya como fuente de información semántica, a modo de espejo para detectar equivalencias (Dyvik, 2004). Esto es, dos elementos en una lengua se pueden considerar similares entre sí cuando el CP revela que tienen los mismos equivalentes en la otra lengua. En el ámbito de los estudios vinculados a expresiones modales ya se han realizado estudios utilizando este recurso, como, por ejemplo, el de Almeida & Carrió Pastor (2015), aunque con un enfoque cualitativo. Desde el enfoque cuantitativo, están relacionados con este estudio los trabajos de Robledo & Nazar (2018) y Nazar (2021), que explotan los CP para extraer marcadores discursivos. La metodología en esos casos es, sin embargo, distinta, ya que se basan en técnicas de *clustering*. La propuesta que exponemos en este artículo no requiere este tipo de algoritmos, que suelen ser computacionalmente costosos. En su lugar, proponemos un método comparativamente más simple, lo cual es preferible según el principio de parsimonia.

El CP utilizado es el Opus Corpus (Tiedemann, 2012), en particular, el subcorpus Scielo de esta colección. Este subconjunto está compuesto por títulos y resúmenes de artículos de investigación de la base de datos del mismo nombre[3]. Esta muestra tiene una extensión de 25.106.776 tokens y fue elegido por corresponder al género argumentativo, terreno fértil para la producción de OM. En el Cuadro 1 se muestra un ejemplo de organización del CP Scielo a partir de la búsqueda del adverbio modal *claramente*.

Como parte de la preparación previa de este material, por cuestión de eficiencia computacional, convertimos el CP del formato TMX original (Figura 1) a un formato TXT en el que en una misma línea se disponen las concordancias alineadas, separadas por un tabulador.

[3] https://scielo.org

3.2. Fases del procedimiento

3.2.1. Creación de un listado inicial de OM en castellano

Se creó un listado inicial de ejemplos de OM del castellano, elaborado inicialmente a partir de los datos de los proyectos Dismark[4] y Text·a·Gram,[5] y aumentado por medio de la anotación manual de un corpus de columnas de opinión, siguiendo los lineamientos de Nissim et al. (2013).

El resultado de esta fase cualitativa inicial fue un listado de 93 OM aléticos (por ejemplo, *podría ser que, con toda probabilidad, es esperable*), 236 epistémicos (*indudablemente, nos consta, claro está*), 142 deónticos (*es fundamental que, necesariamente, urge*) y 188 valorativos (*importantísimo, es favorable, es muy inapropiado*). Nos referiremos a este listado como el conjunto E_m.

3.2.2. Extracción de n-gramas del CP

En cualquiera de las lenguas con las que se trabaje, el material de entrada o input consiste en un listado de vocabulario. En este caso, naturalmente, ese vocabulario procede de una de las lenguas del CP. Se ordenó el vocabulario del CP en listados palabras y secuencias de hasta cinco palabras, definiendo así un conjunto V de n-gramas con $n \leq 5$. V es entonces el conjunto de unidades input ($x \in V$).

Como paso previo del proceso de clasificación aplicamos a V un filtro por medio de un etiquetador morfológico — UDPipe (Straka & Straková, 2017) — para descartar los n-gramas que inician con sustantivo y aquellos que contienen formas verbales con pronombres enclíticos (*promoverse, pautearse, sugerirse*, etc.), ya que son características que no se asocian con los OM.

Un segundo filtro consiste en retener solamente las unidades más frecuentes: las primeras 100.000 en el caso de las monoléxicas, y las primeras 25.000 en el caso de las secuencias de palabras.

3.2.3. Clasificación

Cada elemento x del conjunto V es tomado como input para una función que devuelve un valor binario (*True/False*) para la proposición $x \in OM$. Esta función se basa en la medición de la coocurrencia en el CP entre x y cualquier miembro del ejemplario E_m. Definimos para ello

[4] http://www.tecling.com/dismark
[5] http://www.tecling.com/textagram

	Texto en castellano	**Texto en inglés**
1	Yo percibo eso claramente en el día a día.	I see it clearly in my daily routine.
2	El título lo decía claramente: Arquitectura y negocios.	The title made it clear: Arquitectura y negocios Architecture and Business.
3	Esto se puede observar claramente en la Ilustración 2.	This can clearly be seen in figure 2.

Cuadro 1: Ejemplo de estructura del CP Scielo.

```
    <tu>
      <tuv xml:lang="en"><seg>This is obviously a definition
that incorporates, in this case, sufficient scientific nuance
and, from a lexicographical point of view, minimizes the effect
of circularity by omitting the term dropsy, equivalent to an
accumulation of serous fluid above typical levels.</seg></tuv>
      <tuv xml:lang="es"><seg>Se trata, evidentemente, de una
definición que incorpora, en este caso, matices científicos
suficientes y, desde el punto de vista lexicográfico, minimiza
los efectos de la circularidad al suprimir el término hidropesía
que equivale a una
acumulación de líquido seroso por encima de los niveles
típicos.</seg></tuv>
    </tu>
    <tu>
```

Figura 1: Muestra de segmentos alineados del corpus paralelo en formato TMX.

un conjunto $R(x)$ como el subconjunto de concordancias de la expresión ingresada como input (x) en el CP (1). A partir de la intersección de $R(x)$ con E_m en el CP (2), obtenemos una función $mop(x)$ (3) que mide la coaparición del input x con algún elemento del conjunto E_m en el CP. La decisión se toma por medio de un umbral arbitrario k (4).

$$R(x) = x \cap CP \tag{1}$$

$$int(x) = |R(x) \cap E_m| \tag{2}$$

$$mop(x) = \frac{int(x)}{|R(x)|} \tag{3}$$

$$\forall x \in V, (x \in OM) = \begin{cases} \text{True} & if\ mop(x) > k \\ \text{False} & otherwise \end{cases} \tag{4}$$

En el Cuadro 2 ejemplificamos el proceso en el caso del castellano como lengua objetivo (y, por tanto, con un E_m en inglés) y con x = *claramente*. En la fila 1, la concordancia del CP en inglés no contiene ningún elemento del conjunto E_m (e.g., *predominantly* $\notin E_m$). En la fila 2, en cambio, sí se presenta un caso de intersección de $R(x)$ con elementos del conjunto E_m (*clearly* $\in E_m$). De este modo, el número de veces en que *claramente* aparece en paralelo con algún OM del ejemplario en inglés (E_m) se divide por la cantidad total de ocasiones en que *claramente* aparece en el CP ($R(x)$). Mientras más alta sea esta proporción, mayor la probabilidad de que $x \in OM$.

3.2.4. Repetición del proceso en orden inverso

Si la lengua objetivo es castellano, en el resultado de la primera ejecución es un listado de OM en inglés como, por ejemplo en el caso de epistémicos, $E_m = \{$*certainly, I believe, undoubtedly, ...*$\}$, el paso siguiente consiste en repetir el mismo proceso a la inversa, es decir, utilizar este resultado intermedio en inglés para obtener un conjunto de OM de vuelta al castellano.

4. Resultados

A continuación se describen los resultados de la aplicación de la propuesta metodológica para las categorías de OM aléticos, epistémicos, deónticos y valorativos. Primero presentamos resultados de la aplicación desde un ejemplario del castellano, es decir que los resultados intermedios están en inglés. Posteriormente, presentamos resultados de la segunda aplicación, en el que se utiliza el ejemplario en inglés ahora para obtener el resultado final en castellano. Aunque la investigación considera n-gramas con $n \leq 5$, por limitaciones de espacio solo presentamos algunos ejemplos de tablas de resultados con bigramas y trigramas. También resumimos en gráficas la evaluación del desempeño del algoritmo de detección de OM del castellano en los 5 n-gramas que re-

	CP_o **(castellano)**	CP_m **(inglés)**
1	La producción obtenida se reveló **claramente** internacional, con apenas un trabajo producido en Brasil.	The obtained production was **predominantly** international, with only one study produced in Brazil.
2	Definir **claramente** la cuestión a plantearse.	Define **clearly** the question to be formulated.

Cuadro 2: Ejemplos de no coincidencia (fila 1) y de coincidencia (fila 2) entre OM de ambas lenguas.

N°	**x**	**R(x)**	$int(x)$	$mop(x)$
1	probably related	37	34	89
2	possibly due	67	60	88
3	possibly because	44	39	86
4	possible that	293	251	85
5	probably due	174	149	85
6	generally associated	27	24	85
7	likely that	109	93	84
8	probably because	75	64	84
9	will probably	33	28	82
10	are probably	60	49	80
11	probable that	52	42	79
12	and probably	49	38	76
13	may mean	29	23	76
14	usually occurs	31	24	75
15	and possibly	58	44	74
16	commonly associated	34	26	74
17	this can	384	276	71
18	probably the	73	53	71
19	usually present	27	20	71
20	was probably	46	33	70

Cuadro 3: Muestra de las más altas puntuaciones obtenidas de bigramas en función de la búsqueda de OM aléticos en inglés

N°	**x**	**R(x)**	$int(x)$	$mop(x)$
1	must be the	39	38	95
2	is necessary that	95	91	94
3	are not necessarily	38	37	94
4	necessary that the	42	40	93
5	*the nurse must	26	25	92
6	should be carefully	35	33	91
7	care should be	49	45	90
8	is not necessarily	43	40	90
9	*the physician should	30	28	90
10	must be able	29	27	90
11	necessary to consider	76	69	89
12	there must be	66	60	89
13	professionals should be	46	42	89
14	must be based	46	42	89
15	necessary to know	38	35	89
16	*patients must be	36	33	89
17	*should be informed	28	26	89
18	should be avoided	96	86	88
19	must consider that	26	24	88
20	we must be	25	23	88

Cuadro 4: Muestra de las más altas puntuaciones obtenidas de trigramas en función de la búsqueda de OM deónticos en inglés

presentan nuestros resultados finales. El resto de los datos están en la ya mencionada web del proyecto.

4.1. Resultados intermedios en inglés

A continuación presentamos una muestra con los resultados de los 20 puntajes $mop(x)$ más altos de la aplicación del procedimiento a un listado de bigramas y trigramas en inglés. Los bigramas corresponden, en este caso, a OM aléticos (Cuadro 3) y los trigramas a los OM deónticos (Cuadro 4). Se marcan con asterisco los casos de elementos que no corresponden a un OM. El número total de unidades en inglés es de 512 en el caso de los aléticos, 338 en el caso de los epistémicos, 469 en el caso de los deónticos y 684 en el caso de los valorativos.

4.2. Resultados finales en castellano

La aplicación del procedimiento con V ahora compuesto por listados de n-gramas en castellano con $n \leq 5$ resultó en un total de 1.084 casos de OM (Cuadro 5). Como en el caso anterior, presentamos una muestra con los resultados finales de los 20 puntajes más altos de la aplicación del procedimiento a n-gramas en castellano. El Cuadro 6 presenta bigramas con OM aléticos y el Cuadro 7 con epistémicos. Al igual que en los casos anteriores, se marcan con asterisco los n-gramas mal clasificados.

Para mostrar la evaluación del desempeño del algoritmo de detección de OM en estos resultados finales en castellano, desde las Figuras 2 a la 5 presentamos gráficos de líneas con la precisión acumulada en los primeros 100 candidatos a OM aléticos (Figura 2), epistémicos (Figura 3), deónticos (Figura 4) y valorativos (Figura 5), en

Categoría	**Cantidad**
Aléticos	69
Epistémicos	160
Deónticos	653
Valorativos	202
Total	**1084**

Cuadro 5: Distribución de OM detectados en castellano por categoría

N°	**x**	**R(x)**	$int(x)$	$mop(x)$
1	probablemente debido	51	49	94
2	posiblemente debido	30	29	93
3	probablemente porque	26	25	92
4	pueda causar	9	9	90
5	pudiendo causar	16	15	88
6	nunca será	8	8	88
7	debido posiblemente	7	7	87
8	podría explicarse	28	25	86
9	posible explicación	56	48	84
10	*variar desde	12	11	84
11	podría atribuirse	12	11	84
12	pudiendo llevar	23	20	83
13	*sólo podrá	11	10	83
14	posiblemente porque	16	14	82
15	*explicarse porque	10	9	81
16	* ocurrir después	10	9	81
17	*causar cambios	10	9	81
18	*ocurrir durante	29	24	80
19	pudiendo incluso	9	8	80
20	pudiendo resultar	13	11	78

Cuadro 6: Muestra de las más altas puntuaciones obtenidas de bigramas en función de la búsqueda de OM aléticos en castellano

los cinco tipos de n-gramas ($1 \leq n \leq 5$). El patrón que se observa es que muchos OM se detectan correctamente al inicio y luego la precisión comienza gradualmente a decaer.

4.3. Métricas de evaluación

La precisión de estos resultados finales en castellano fue evaluada cualitativamente por dos anotadores. Para medir el grado de acuerdo se evaluó una muestra aleatoria de 100 casos por cada una de las categorías, constituidas por los distintos tipos de n-gramas en partes iguales. Esta medición arrojó un acuerdo total del 94 % y un índice Kappa de Cohen de 0.89, que puede considerarse alto. La evaluación de la precisión de los resultados fue medida a partir de muestras de 100 casos por cada categoría de OM estudiada, 400 en total. Se seleccionaron muestras de

N°	**x**	**R(x)**	$int(x)$	$mop(x)$
1	aparentemente sanos	18	18	94
2	yo pienso	163	148	90
3	ninguna duda	10	10	90
4	piensa usted	10	10	90
5	claramente definido	9	9	90
6	probablemente porque	26	24	88
7	entonces pienso	8	8	88
8	debido posiblemente	7	7	87
9	lógicamente estabilizado	7	7	87
10	probablemente debido	51	45	86
11	posiblemente debido	30	26	83
12	*usted cree	5	5	83
13	yo creo	200	165	82
14	posiblemente porque	16	14	82
15	queda claro	64	52	80
16	nuestra opinión	53	42	77
17	posible pensar	21	17	77
18	tengo dudas	8	7	77
19	quizá porque	8	7	77
20	resulta claro	8	7	77

Cuadro 7: Muestra de las más altas puntuaciones obtenidas de bigramas en función de la búsqueda de OM epistémicos en castellano

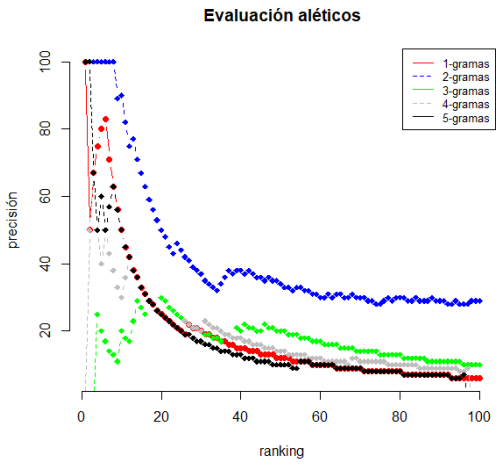

Figura 2: Resultados de la detección de OM aléticos con n-gramas ($1 \leq n \leq 5$)

n-gramas del mismo tamaño a partir de distintas bandas de frecuencia y según la puntuación obtenida: alta, baja y media. Los resultados de este procedimiento se sistematizan en el Cuadro 8. Tal como se puede observar, la mejor precisión del algoritmo se logra en las categorías deóntica (98 %) y epistémica (95 %). En cuanto a los tipos de n-gramas, la mayor precisión del algoritmo se logra con $n = 1$ (93.7 %) y $n = 2$ (96.2 %). El análisis de estos resultados se explicita en la sección 4.4.

n-gramas	aléticos	epistémicos	deónticos	valorativos	T. por n-grama
n = 1	80 %	95 %	100 %	100 %	**93.7 %**
n = 2	95 %	100 %	100 %	90 %	**96.2 %**
n = 3	60 %	95 %	100 %	80 %	**83.7 %**
n = 4	70 %	95 %	100 %	75 %	**85 %**
n = 5	80 %	90 %	90 %	55 %	**78.7 %**
Total	**77 %**	**95 %**	**98 %**	**80 %**	

Cuadro 8: Evaluación de la precisión de detección del algoritmo por categorías de OM y por n-gramas

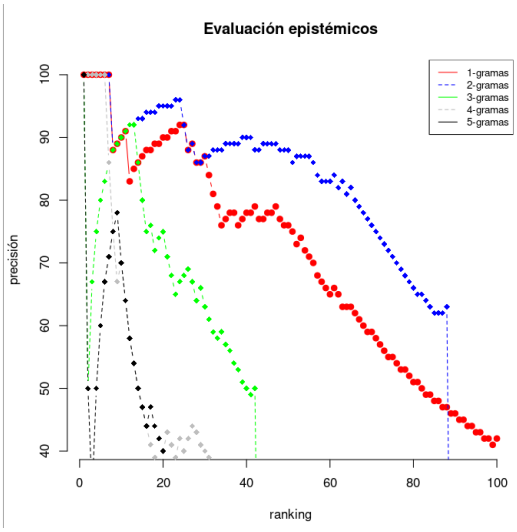

Figura 3: Resultados de la detección de OM epistémicos con n-gramas ($1 \leq n \leq 5$)

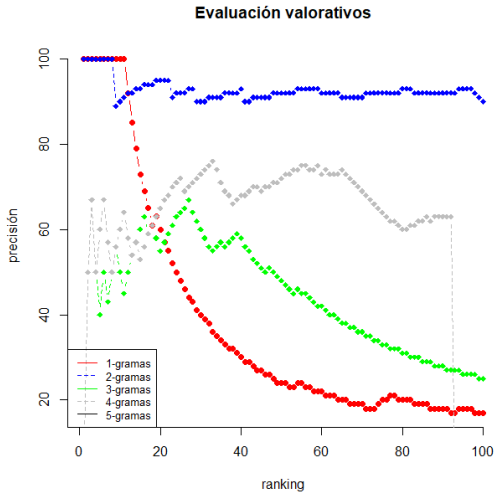

Figura 5: Resultados de la detección de OM valorativos con n-gramas ($1 \leq n \leq 5$)

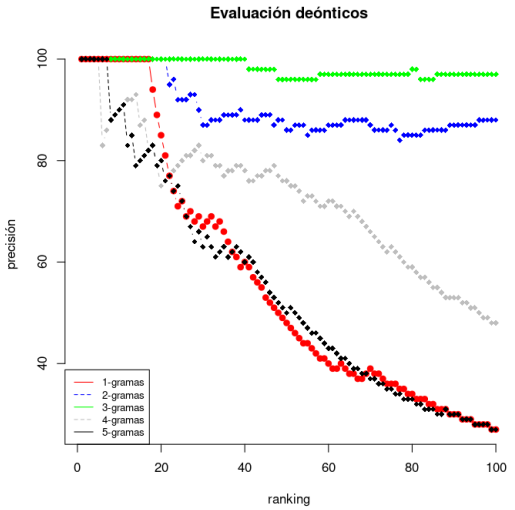

Figura 4: Resultados de la detección de OM deónticos con n-gramas ($1 \leq n \leq 5$)

Como una forma de evaluar la cobertura del método nos inspiramos en el trabajo de Lopes et al. (2015) para desarrollar un *baseline* o método de base. Estos autores se interesan por los marcadores discursivos y aplican un traductor automático para aumentar un listado inicial elaborado manualmente. Para nuestro *baseline*, entonces, imitamos el procedimiento utilizando un traductor automático (DeepL[6]) para convertir el ejemplario inicial en castellano a uno en inglés y utilizar el resultado para traducir de nuevo al castellano. Con este fin, dispusimos los OM del ejemplario inicial en un listado, es decir, uno por línea y seguidos de punto.

El Cuadro 9 muestra los resultados del método propuesto y del *baseline*, y se percibe una diferencia bien marcada. Como se puede observar, en el caso del método propuesto se obtiene mayor diversidad de OM. El traductor automático ofrece sistemáticamente menor variedad. Contrariamente a lo esperado, la intersección entre los resultados obtenidos con nuestro método y los del *baseline* es baja, a tal punto que hace pensar en la posibilidad de combinar ambos métodos en el futuro. En cualquier caso, la intersección con el ejemplario que se muestra en las últimas dos columnas para cada método resume la diferencia en crecimiento con respecto al ejemplario inicial. En todos los casos la comparación favorece a nuestro método por amplio margen, ya que el sistema proporciona más OM y menos coincidencia

[6] https://www.deepl.com/translator

con el ejemplario en comparación con el *baseline*, que con mayor frecuencia acaba reproduciendo los OM del ejemplario inicial.

Otra medida para determinar la cobertura de los resultados obtenidos fue contrastarlos con un listado de 167 OM registrados en el *Diccionario de conectores y operadores del español* (Fuentes Rodríguez, 2009). Esta fuente se eligió porque representa el registro más reciente de estas unidades en castellano. Esta evaluación consideró dos acciones: 1) determinar cuáles de los 167 OM registrados por Fuentes Rodríguez (2009) se encuentran también en el subcorpus Scielo del Opus Corpus y 2) comparar la intersección resultante con los listados obtenidos en esta investigación.

Como resultado de la primera acción se obtuvo que 117 OM registrados en el diccionario de Fuentes Rodríguez (2009) se encuentran también en el corpus Scielo. Como resultado del segundo paso se obtuvo que 15 OM de ese listado fueron detectados por el algoritmo desarrollado en esta aplicación del procedimiento. De acuerdo con esta evaluación, la medida de la cobertura es del 13 %, un número evidentemente bajo. Hay por lo menos dos aspectos para tener en cuenta en la interpretación de este dato.

En primer lugar, los OM registrados por Fuentes Rodríguez (2009) no discriminan entre los que son utilizados en la lengua escrita y en la oral (*a ver, oye, 'eso, eso'*). Aunque algunas de estas unidades pueden encontrarse en un corpus escrito como el de Scielo, su aparición en la escritura no necesariamente está asociada a la expresión de un componente modal de los enunciados, tal como sucedería en el caso de la modalidad oral y, por este motivo, no sería parte de nuestro objetivo detectarlas. En segundo lugar, aunque la intersección de ambos listados es baja, el resultado del registro total de candidatos a OM de las cuatro categorías estudiadas a través de esta propuesta metodológica (1.084 casos) supera ampliamente a los 167 casos registrados por la autora. En esta línea, cabe decir que la medición que hicimos revela que, aun cuando aquí dispongamos de una gran cantidad de OM, debemos concluir que realmente existen muchos más casos por detectar. En este sentido, utilizar CP con otras características, es decir, que incorporen otros géneros discursivos, es una tarea de futuro necesaria para ampliar el registro hasta ahora obtenido.

4.4. Análisis de los resultados

En función de los resultados y su evaluación, relevamos los siguientes aspectos.

En primer lugar, considerando el resumen detallado en el Cuadro 8, se comprueba que la mayor precisión es obtenida en las categorías de OM deónticos y epistémicos. Este resultado podría estar relacionado con que son dos tipos de modalización cuya manifestación lingüística es muy fuerte y, por lo tanto, siempre es muy marcada, incluso dentro de los discursos académicos como los que constituyen el subcorpus Scielo. Este hecho viene a reafirmar la conveniencia de estudiar el fenómeno de la modalización a partir de la división epistémico / deóntica, en tanto categorías modalizadoras fuertes (Müller, 2007; Portner, 2009).

La caída en la precisión de los OM valorativos presentada en el Cuadro 8 desde los n-gramas de 1 y 5 palabras (100 % y 55 %, respectivamente), consideramos que puede estar relacionada con la falta de heterogeneidad semántica, y no solo formal, dentro del ejemplario inicial del proceso (3.2.1). Otra variable de peso podría ser el hecho de tratarse de un CP de lenguaje académico, donde –por estilo y norma– la manifestación de la valoración debe ser restringida. Como resultado preliminar es positivo, pero futuras aplicaciones deberían considerar la posibilidad de nutrir el ejemplario inicial con unidades léxicas provenientes del análisis de sentimientos (Liu, 2010; Zhang et al., 2018, por ejemplo). Además, es probable que a medida que la secuencia de palabras vaya aumentando, su componente modalizador vaya perdiendo fuerza. Estos mismos fenómenos, considerando que no constituyen una marca de manifestación modal fuerte, podrían estar afectando la obtención de una medida de precisión más estable en el caso de los OM aléticos.

Otro aspecto a considerar sería la posibilidad de excluir del análisis las secuencias de más de tres palabras. Esto es porque los resultados muestran que que el *peak* estable de precisión se obtiene en los bigramas (Cuadro 8) y hay motivos para sospechar que el componente modalizador va perdiendo fuerza con el aumento de tamaño de la secuencia de palabras.

Finalmente, a diferencia de los registros existentes, los resultados obtenidos a partir de esta metodología evidencian un aspecto importante sobre la naturaleza de los OM: que no se corresponden ni única ni principalmente con expresiones o construcciones lingüísticas con alto grado de gramaticalización o estabilidad. En esa línea, será necesario un análisis de la estructura formal los OM identificados y de su variación dentro de un mismo paradigma, metodología que excede, por supuesto, los límites de lo planteado en este artículo.

OM	E	M	B	$\|M \cap B\|$	$\|E \cap M\|$	$\|E \cap B\|$
Epistémico	236	160	150	11	23	90
Deóntico	142	653	77	7	13	49
Alético	93	69	71	4	5	51
Valorativo	188	202	152	3	7	112

Cuadro 9: Comparación con un *baseline* (E = ejemplario; M = método propuesto y B = *baseline*)

5. Conclusiones y trabajo futuro

Se ha presentado una propuesta metodológica para detectar OM a partir de un CP con un método principalmente estadístico. En este caso, se han presentado resultados intermedios en inglés que luego fueron utilizados como insumo para la obtención de resultados finales en castellano.

Respecto al desarrollo del algoritmo y los resultados que presentamos, observamos en general un desempeño aceptable, sobre todo en la identificación de OM epistémicos y deónticos.

Una de las limitaciones del estudio es que no podemos ofrecer una comparación con otros métodos más allá del *baseline*, ya que esta es, que sepamos, la primera vez que se propone una tarea de identificación y registro de OM con métodos del procesamiento de lenguaje natural. Es de esperar que futuras propuestas mejoren esta primera aproximación al problema. Otra limitación es que, aunque se ha evaluado la cobertura del método a partir de los OM proporcionados por un diccionario (4.3), sostenemos que esta evaluación pudiera ser parcial u objetable porque no existen, por ahora, registros de OM suficientemente completos en castellano que puedan utilizarse como referencia. En ese sentido, la constitución de un listado de contraste a partir de la anotación manual de un corpus extenso podría ser otra posibilidad que dejamos también para el futuro.

Quedará también para el futuro mejorar el desempeño de este clasificador, en particular cuando se trabaja con n-gramas de $n > 3$. Una posibilidad para ello sería detectar y eliminar palabras en otras lenguas (por ejemplo, *necessariamente*, en portugués) o términos de dominio especializado (por ejemplo, *cianogénicas* o *monoinsaturados*) que a veces se seleccionan por error debido a la naturaleza especializada del subcorpus Scielo. Para ello podría servir un extractor terminológico. También se podría experimentar con CP de otras características, como por ejemplo uno que tenga mayor riqueza expresiva que el discurso académico. Esto, a su vez, redundaría en la obtención de una mayor variedad de OM.

Otros desafíos interesantes para el futuro serán reproducir experimentos con otras lenguas e incluso estudiar los préstamos de modalizadores entre lenguas, fenómeno que ha sido parcialmente explorado por der Auwera & Ammann (2005). También sería interesante estudiar el grado de modulación (alta, media o baja) de un OM, y si existe alguna relación entre este grado y la medida $mop(x)$ de esta propuesta.

Por último, considerando que las categorías analizadas aquí no agotan las posibilidades de estudios de la expresión de la modalización (Narrog, 2012), es parte del trabajo en curso seguir explorando esta metodología con otras categorías (veredictorias, volitivas, de usualidad, entre otras). Ello significará un aporte útil además para evaluar cuáles OM pueden ser polifuncionales entre distintas categorías.

Agradecimientos

El primer autor agradece el apoyo financiero de la Beca de Magíster Nacional/2021 de la Agencia Nacional de Investigación y Desarrollo (ANID) del Gobierno de Chile, que permitió desarrollar esta investigación. Agradecemos también a las revisoras por su trabajo.

Referencias

Almeida, Francisco A. & María. L. Carrió Pastor. 2015. Sobre la categorización de seem en inglés y su traducción en español. análisis de un corpus paralelo. *Revista Signos* 48(88). 154–173. 10.4067/S0718-09342015000200001.

der Auwera, Johan Van & Andreas Ammann. 2005. Modal polyfunctionality and standard average european. En *Modality: Studies in Form and Function*, 247–272. Equinox.

der Auwera, Johan Van & Vladimir A. Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2. 79–124. 10.1515/lity.1998.2.1.79.

Baker, KKathryn, Bloodgood Michael, Bonnie J. Dorr, Nathaniel W. Filardo, Lori Levin &

Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. En *7th International Conference on Language Resources and Evaluation (LREC)*, 1402–1407.

Barrenechea, Ana María. 1979. Operadores pragmáticos de actitud oracional: los adverbios en mente y otros signos. *Estudios lingüísticos y dialectológicos* 39–59.

Benveniste, Émile. 1966. *Problemas de lingüística general I.* Siglo XXI.

Benveniste, Émile. 1974. *Problemas de lingüística general II.* Siglo XXI.

Bernárdez, Enrique. 1982. *Introducción a la lingüística del texto.* Espasa-Calpe.

Blanché, Robert. 1966. *Structures intellectuelles.* Vrin Reprise.

Brandt, Søren. 1999. *Modal verbs in Danish.* C.A. Reitzel.

Calsamiglia, Helena & Amparo Tusón. 1999. *Las cosas del decir: manual de análisis del discurso.* Ariel.

Casado Velarde, Manuel. 1993. *Introducción a la gramática del texto en español.* Arco Libros.

Charaudeau, Patrick. 1994. *Grammaire du sens et de l'expression.* Hachette.

Cuenca, María Josep. 2010. *Gramática del texto.* Arco/Libros.

De Beaugrande, Robert A. & Wolfgang. U. Dressler. 1997. *Introducción a la lingüística del texto.* Ariel.

Dyvik, Helge. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. En *23rdrd International Conference on English Language Research on Computerized Corpora (ICAME)*, 309–326. 10.1163/9789004333710_019.

Fuentes Rodríguez, C. 2009. *Diccionario de conectores y operadores del español.* Arco/Libros.

Fuentes Rodríguez, Catalina. 2003. Operador/conector, un criterio para la sintaxis discursiva. *Rilce* 19(1). 61–85. 10.15581/008.19.26730.

Ghia, Elisa, Lennart Kloppenburg, Malvina Nissim, Paola Pietrandrea & Valerio Cervoni. 2016. A construction-centered approach to the annotation of modality. En *12th ISO Workshop on Interoperable Semantic Annotation*, 67–74.

González, Ramón, Dámaso Izquierdo & Óscar Loureda. 2016. *La evidencialidad en español: teoría y descripción.* Iberoamericana Vervuert. 10.31819/9783954878710.

Greimas, Algirdas J. 1973. Les actants, les acteurs et les figures. En *Sémiotique narrative et textuelle*, 161–176. Larousse.

Gutiérrez, Rosa M. 2010. Especialización del discurso: Una caracterización desde el sistema de la obligación. *Revista de Lingüística Teórica y Aplicada* 48(1). 105–132. 10.4067/S0718-48832010000100006.

Hendrickx, Iris, Amália Mendes & Silvia Mencarelli. 2012. Modality in text: a proposal for corpus annotation. En *8th International Conference on Language Resources and Evaluation (LREC)*, 1805–1812.

Kalinowski, Georges. 1976. Un aperçu élémentaire des modalités déontiques. *Langages* 43. 10–18.

Kerbrat-Orecchioni, Catherine. 1987. *La enunciación: de la subjetividad en el lenguaje.* Edicial.

Lee, Kenton, Yoav Artzi, Yejin Choi & Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1643–1648. 10.18653/v1/D15-1189.

Liu, Bing. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing* 627–666.

Lopes, António, David Martins, Vera Cabarrão, Ricardo Ribeiro, Helena Moniz, Isabel Trancoso & Ana Isabel Mata. 2015. Towards using machine translation techniques to induce multilingual lexica of discourse markers. ArXiv [cs.CL]. 10.48550/arXiv.1503.09144.

Lozano, Jorge, Cristina Peña-Marín & Gonzalo Abril. 1982. *Análisis del discurso: Hacia una semiótica de la interacción textual.* Cátedra.

Marasović, A., M. Zhou, A. Palmer & A. Frank. 2016. Modal sense classification at large: Paraphrase-driven sense projection, semantically enriched classification models and cross-genre evaluations. *Linguistic Issues in Language Technology (LiLT)* 14. 1–58.

Martín-Zorraquino, M. Antónia & José Portolés. 1999. Los marcadores del discurso. En *Gramática descriptiva de la lengua española*, vol. 3, 4051–4213. Espasa. 10.15581/008.16.27325.

Miwa, Makoto, Paul Thompson, John McNaught, Douglas B. Kell & Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical

literature. *BMC Bioinformatics* 13. 108. 10.1186/1471-2105-13-108.

Müller, Gisela. 2007. Metadiscurso y perspectiva: Funciones metadiscursivas de los modificadores de modalidad introducidos por 'como' en el discurso científico. *Revista Signos* 40(64). 357–387. 10.4067/S0718-09342007000200005.

Narrog, Heiko. 2012. *Modality, subjectivity, and semantic change: A cross-linguistic perspective*. Oxford University Press.

Nazar, Rogelio. 2021. Inducción automática de una taxonomía multilingüe de marcadores discursivos: primeros resultados en castellano, inglés, francés, alemán y catalán. *Procesamiento del Lenguaje Natural* 67. 127–138.

Nissim, Malvina, Paola Pietrandrea, Andrea Sansó & Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. En *9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 7–14.

Nuyts, Jan. 2005. The modal confusion: on terminology and the concepts behind it. En *Modality: Studies in Form and Function*, 5–38. Equinox.

Nuyts, Jan. 2006. Modality: Overview and linguistic issues. En *The Expression of Modality*, 1–26. De Gruyter. 10.1515/9783110197570.1.

Nuyts, Jan. 2016a. Analyses of de modal meanings. En *The Handbook of Modality and Mood*, 31–49. Oxford University Press.

Nuyts, Jan. 2016b. Surveying modality and mood: An introduction. En *The Handbook of Modality and Mood*, 1–8. Oxford University Press.

Otaola, Concepción. 1988. La modalidad (con especial referencia a la lengua española). *Revista de Filología Española* 68(1/2). 97–117. 10.3989/rfe.1988.v68.i1/2.414.

Palmer, Frank R. 2001. *Mood and modality*. Cambridge University Press. 10.1017/CBO9781139167178.

Pérez Canales, José. 2009. *Marcadores de modalidad epistémica: un estudio contrastivo (francés-español)*: Universitat de València. Tesis Doctoral.

Portner, Paul. 2009. *Modality*. Oxford University Press.

Pottier, Bernard. 1977. *Lingüística general: teoría y descripción*. Gredos.

Pyatkin, Valentina, Shoval Sadde, Aynat Rubinstein, Paul Portner & Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. ArXiv [cs.CL]. 10.48550/arXiv.2106.08037.

Quaresma, Paulo, Amália Mendes, Iris Hendrickx & Teresa Gonçalves. 2014. Automatic tagging of modality: identifying triggers and modal values. *Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation* 95–101.

Ridruejo, Emilio. 1999. Modo y modalidad. El modo en las subordinadas sustantivas. En *Gramática descriptiva de la lengua española*, vol. 2, 3209–3252. Espasa.

Robledo, Hermán & Rogelio Nazar. 2018. Clasificación automatizada de marcadores discursivos. *Procesamiento del Lenguaje Natural* 61. 109–116. 10.26342/2018-61-12.

Rubinstein, Aynat, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz & Paul Portner. 2013. Toward fine-grained annotation of modality in text. En *IWCS workshop on annotation of modal meanings in natural language*, 38–46.

Ruppenhofer, Josef & Ines Rehbein. 2012. Yes we can!? Annotating English modal verbs. En *8th International Conference on Language Resources and Evaluation (LREC)*, 1538–1545.

Salas, Millaray. 2015. Una propuesta de taxonomía de marcadores metadiscursivos para el discurso académico-científico escrito en español. *Revista Signos* 48(87). 95–120. 10.4067/S0718-09342015000100005.

Santos Río, Luis. 2003. *Diccionario de partículas*. Luso-Española de Ediciones.

Saurí, Roser, Robert Knippen, Marc Verhagen & James Pustejovsky. 2005. Evita: a robust event recognizer for QA systems. En *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 700–707.

Saurí, Roser, Marc Verhagen & James Pustejovsky. 2006. Annotating and recognizing event modality in text. En *19th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 333–338.

Sologuren, Enrique & René Venegas. 2022. Marcadores epistémicos en el género trabajo final de grado en español: variación disciplinar en la escritura de formación académica. *Literatura y lingüística* 45. 235–258. 10.29344/0717621X.45.2200.

Soni, Sandeep, Tanushree Mitra, Eric Gilbert & Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. En *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 415–420. 10.3115/v1/P14-2068.

Straka, M. & J. Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. En *CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. 10.18653/v1/K17-3009.

Taboada, Maite. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2. 325–347. 10.1146/annurev-linguistics-011415-040518.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. En *8th International Conference on Language Resources and Evaluation (LREC)*, 2214–2218.

Van Dijk, Teun A. 1980. *Texto y contexto: Semántica y pragmática del discurso*. Cátedra.

Van Dijk, Teun A. 2012. *Discurso y contexto*. Gedisa.

Wiebe, Janyce, Theresa Wilson & Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39(2). 165–210. 10.1007/s10579-005-7880-9.

Zhang, Lei, Shuai Wang & Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(e1253). 10.1002/widm.1253.

Proposta recebida em Agosto 2023 e aceite para publicação em Outubro 2023.

Recursos linguísticos para o PLN específico de domínio: o Petrolês

NLP resources for the oil & gas domain: Petrolês

Cláudia Freitas
PUC-Rio

Elvis de Souza
PUC-Rio

Maria Clara Castro
PUC-Rio

Tatiana Cavalcanti
PUC-Rio

Patricia Ferreira da Silva
Petrobras/CENPES

Fábio Corrêa Cordeiro
Petrobras/CENPES
FGV/EMAp

Resumo

Muitas organizações têm dificuldade em recuperar e extrair informações dos seus repositórios de documentos técnicos, em especial operadoras de óleo e gás que há várias décadas acumulam relatórios e documentos geocientíficos. No entanto, a maior parte dos recursos linguísticos para o processamento de linguagem natural é extraída de páginas da internet em inglês. Neste artigo, apresentamos os recursos linguísticos desenvolvidos ao longo do projeto Petrolês, com ênfase no PetroNer, *corpus* padrão ouro anotado com entidades do domínio, dependências sintáticas, e alinhado a uma ontologia de conceitos geológicos. Relatamos o processo de construção do PetroGold, *treebank* padrão ouro usado na geração de um modelo customizado para anotação de dependências sintáticas, e detalhamos o processo de anotação de entidades no PetroNer, realizado por meio de regras. Também realizamos um estudo sobre a aplicação das regras no *corpus* e, por fim, descrevemos características linguísticas do material que compõe o Petrolês, comparando-o com um *corpus* de textos jornalísticos.

Palavras chave

entidades mencionadas, ontologia geológica, dependências sintáticas, *universal dependencies*, *corpus* padrão ouro

Abstract

Many organizations struggle with retrieving and extracting information from their repositories of technical documents, particularly oil and gas operators with decades of accumulated geoscientific reports and documents. However, the majority of linguistic resources for natural language processing are derived from internet pages in English. In this article, we present the linguistic resources developed throughout the Petrolês project, with an emphasis on PetroNer, a gold standard corpus annotated with domain entities, syntactic dependencies, and aligned with an ontology of geological concepts. We report the construction process of PetroGold, a gold standard treebank used in generating a customized model for syntactic dependency annotation, and we detail the entity annotation process in PetroNer, carried out through the creation of linguistic rules. We also conduct a study on the application of rules in the corpus, and finally, we describe linguistic characteristics of the material comprising Petrolês, comparing it with a corpus of journalistic texts.

Keywords

named entities, geology ontology, syntactic dependencies, universal dependencies, gold standard portuguese *corpus*

1. Apresentação

Um dos requisitos para um Processamento de Linguagem Natural (PLN) bem-sucedido é a existência de recursos linguísticos de qualidade, capazes de oferecer sustentação para as diversas camadas do processamento automático de textos. Quando a tarefa envolve domínios de especialidade, como a área de petróleo, nosso foco, a necessidade de recursos de qualidade é ainda maior.

Modelos dedicados à resolução de correferência, tarefa e etapa importante na identificação de informação em textos, apresentam queda no desempenho quando aplicados a textos acadêmicos (Cohen et al., 2017). Já na análise sintática, um modelo treinado em um *corpus* composto por textos jornalísticos tem uma queda de mais de 10% em seu desempenho quando utilizado na anotação de um *corpus* do domínio biomédico (Thompson et al., 2017).

Mesmo com a popularização de grandes modelos de linguagem (LLMs), a existência de conjuntos de dados linguísticos de qualidade, pro-

DOI: 10.21814/lm.15.2.412

duzidos originalmente na língua de interesse e específicos para um domínio continuam recursos valiosos. Quanto mais cuidado na preparação dos dados, melhor é a qualidade das predições, com a vantagem de serem necessários menos dados para atingir bons resultados (Souza et al., 2020; Lewkowycz et al., 2022; Samuel et al., 2023).

Petrolês é simultaneamente um *corpus* e um repositório de artefatos de PLN especializados no domínio de petróleo em Português, cujo objetivo é servir como referência para os grupos de pesquisas em inteligência artificial e empresas atuantes nesse domínio. Atualmente, o repositório Petrolês[1] conta com conjuntos de dados linguísticos, como *corpora*, e modelos vetoriais pré-treinados especializados no domínio (Gomes et al., 2021). Neste artigo, apresentamos os *corpora* padrão ouro produzidos e disponibilizados pelo projeto, com especial atenção ao PetroNer, um *corpus* com anotação de entidades do domínio, criado para auxiliar extração de informação em documentos técnicos.

Ao longo de 4 anos de Petrolês foram produzidos o *corpus* Petrolês, composto por documentos acadêmicos (monografias, dissertações e teses), boletins e relatórios técnicos em formato texto simples (Cordeiro, 2020); o PetroTok, um subconjunto padrão ouro no que se refere às etapas de pré-processamento textual, especialmente sentenciação (Cavalcanti et al., 2021); o PetroGold, um *treebank* — *corpus* com anotação sintática — com anotação padrão ouro relativa a dependências sintáticas (de Souza et al., 2021; de Souza, 2023), Petro1 e Petro2, pequenos *treebanks* também padrão ouro, e o PetroNer, que além de dependências sintáticas, contém anotação padrão ouro relativa às entidades de interesse do domínio.

Para a anotação morfossintática, utilizamos o *framework* do projeto *Universal Dependencies* (de Marneffe et al., 2021), e com isso também buscamos alinhar os *treebanks* do Petrolês a *treebanks* de outras línguas, contribuindo para a inserção da língua portuguesa no contexto do processamento multilíngue. A anotação das entidades contou com a supervisão de especialistas da área, e foi realizada por meio da aplicação de regras linguísticas. Todo o procedimento de anotação semântica se inspirou no corte-e-costura (Mota & Santos, 2009; Santos & Mota, 2010), ferramenta de auxílio à anotação semântica dos *corpora* do projeto AC/DC (Santos & Sarmento, 2003; Santos, 2011).

[1] https://petroles.puc-rio.ai/

A construção de *corpora* anotados — em nosso caso, motivada primeiramente pelas demandas do PLN — envolve a tomada de decisões variadas, que incluem a compilação de material que seja representativo daquele para o qual a aplicação foi pensada, e variado com relação aos fenômenos (linguísticos) que contém, consistente com relação às anotações codificadas, bem documentado, e com tamanho que permita treinar e avaliar modelos de aprendizado de máquina ou, pelo menos, avaliar modelos e ferramentas.

No Petrolês, a construção de cada recurso anotado foi acompanhada de algum tipo de estudo experimental. A anotação morfossintática do Petro1, Petro2 e PetroGold possibilitou um estudo sobre métodos de revisão de *treebanks* (Freitas & de Souza, 2023; de Souza, 2023) e um estudo sobre o impacto de representações do conhecimento linguístico no desempenho de modelos de aprendizado de máquina (de Souza & Freitas, 2023). A construção do PetroNer, por sua vez, permitiu um estudo sobre o desempenho de um anotador baseado em regras na anotação de entidades do domínio, apresentado aqui.

Neste artigo, detalhamos as etapas de construção do PetroNer, que partiu de um léxico produzido por especialistas da área, e foi anotado com base em regras linguísticas. Na seção 2 apresentamos *corpora* criados com objetivos semelhantes aos do PetroNer; na seção 3 relatamos brevemente a construção do PetroGold, que foi utilizado como material de treino para o modelo de dependências customizado que anotou o PetroNer. Na seção 4 detalhamos a construção do PetroNer, um *corpus* multicamadas e padrão ouro quanto a entidades mencionadas do domínio. Ainda na seção 4, apresentamos o PetroNer como um *benchmark*, e simulamos o desempenho de um anotador baseado em regras na anotação de entidades. Na seção 5, discorremos sobre as características linguísticas do Petrolês e seu possível impacto no desenvolvimento de modelos de linguagem. Por fim, na seção 6, fazemos nossas considerações finais.

2. Trabalhos relacionados

Dois *corpora* nos serviram de inspiração para o PetroNer: o *corpus* (e projeto) GENIA (Kim et al., 2003; Thompson et al., 2017) e o *corpus* CRAFT (Cohen et al., 2017). O GENIA é um *corpus* da área biomédica, composto por 2 mil resumos/400 mil palavras e quase 100 mil anotações de termos biológicos. Possui uma anotação multicamadas, com POS (*part-of-speech*, ou classes gramaticais), sintaxe, entida-

des, eventos e relações, e foi anotado manualmente. Para a anotação dos termos biomédicos, a equipe do GENIA criou, a partir de *corpus*, sua própria ontologia, e a partir dela foi feita a anotação. Ao longo dos anos, as anotações foram continuamente enriquecidas, fazendo com que o GENIA se tornasse um *corpus* de referência no treinamento e avaliação de diversos sistemas do domínio biomédico.

O CRAFT (Colorado Richly Annotated Full Text) é um *corpus* composto por artigos de biomedicina, com 560 mil *tokens*/21 mil frases, e que também conta com diversas camadas de anotação – morfossintaxe, entidades mencionadas e correferência. Possui as mesmas motivações do GENIA e foi inspirado por ele. No entanto, o CRAFT surgiu da necessidade de se trabalhar com artigos completos, e não apenas com resumos de artigos, após o reconhecimento, segundo os autores, de que o corpo do texto trazia mais informações e que a estrutura textual dos resumos era diferente daquela encontrada nos artigos completos. Do ponto de vista da anotação de entidades biomédicas, a motivação para o esquema de anotação criado pelo CRAFT foi ampliar as classes semânticas utilizadas, circunscritas inicialmente aos genes e produtos a eles relacionados, a fim de possibilitar pesquisas relacionadas a outras classes de entidades.

Exceto pelo CRAFT, boa parte dos *corpora* anotados específicos de domínio, como o *corpus* GENIA, é composta apenas por resumos ou por extratos de parágrafos (por exemplo, Gábor et al. (2018) e Augenstein et al. (2017)).

Para a língua portuguesa, temos os *corpora* de domínios criados por Lopes & Vieira (2013), que foram anotados em formato XML pelo *parser* PALAVRAS, e utilizados sobretudo para estudos relativos à extração de terminologias. Dentre esses *corpora* está o GeoCorpus, um *corpus* com anotação de entidades específico para Bacia Sedimentar Brasileira (Amaral et al., 2017). No entanto, o material tem um escopo de entidades mais restrito, e o processo de anotação não pôde contar com a participação direta e intensa de especialistas. O material possui 5.275 frases e, na versão revista,[2] contém 6.126 anotações de entidades.

Nos *corpora* do Petrolês, especificamente PetroGold e PetroNer, todo conteúdo textual está preservado,[3] e as frases estão disponibilizadas na sequência em que foram escritas. No *treebank* PetroGold, foram excluídos dos documentos apenas informações consideradas irrelevantes para os objetivos da anotação e do projeto, como sumário, agradecimentos, folha de aprovação (no caso de teses e dissertações), lista de siglas e a seção de referências bibliográficas. No PetroNer, composto por boletins e relatórios técnicos, os documentos foram processados integralmente.

Na comparação com seus "pares" GENIA, CRAFT e GeoCorpus, o PetroNer contém 615 mil *tokens*/500 mil palavras e quase 19 mil entidades anotadas/27 mil anotações (uma entidade pode ser composta por mais de um *token*), em um processo de anotação cuidadoso que será descrito em 4.1.

3. Sintaxe e o PetroGold

No PLN, a relevância da análise sintática vai além da sintaxe propriamente, uma vez que este tipo de informação pode auxiliar a extração de relações semânticas, como demonstrado em Nooralahzadeh et al. (2018). Além disso, a atribuição de papeis semânticos, também considerada relevantes na identificação de conteúdo em textos, costuma ser feita a partir de *treebanks* de qualidade (Gildea & Jurafsky, 2000). De um ponto de vista metodológico, a existência de informação morfossintática de qualidade é capaz de otimizar outros tipos de anotação humana, como a anotação de entidades. Poder contar com generalizações linguísticas relativas a elementos coordenados, núcleos e modificadores é de grande ajuda na busca e na revisão semântica, como veremos.

O PetroGold é composto por teses e dissertações, e totaliza cerca de 250 mil *tokens*/9 mil frases. Levando em conta os objetivos do projeto, utilizamos como critério de seleção a presença de palavras candidatas à entidade por documento. A variedade lexical, medida pela distribuição *type/token* por documento, foi usada como critério complementar.

A anotação sintática utilizou a abordagem *Universal Dependencies* (UD) (de Marneffe et al., 2021). Escolhemos UD devido à sua crescente utilização na comunidade PLN, o que traz como benefícios não apenas o desenvolvimento e disponibilização de uma série de ferramentas associadas, como anotadores automáticos e ferramentas de auxílio à anotação e revisão de *treebanks*, mas também a possibilidade de alinhamento a outros *treebanks*, tanto de língua portuguesa quanto de outros idiomas, viabilizando estudos multilíngues e multigêneros.

A construção do *treebank* envolveu diferentes fases, e foi feita a partir da revisão de uma anotação automática. A revisão foi feita

[2] http://github.com/bsconsoli/GeoCorpus-V3

[3] Exceto para os casos de frases com problemas graves de tokenização e sentenciação, que foram excluídas.

inicialmente por 4 anotadores, já familiarizados com a abordagem UD. Na primeira fase de anotação/revisão, alguns documentos foram anotados por todos e as divergências discutidas em grupo, com consulta à documentação do próprio projeto UD. No entanto, a anotação de um *corpus* de domínio e gênero novos trouxe desafios linguísticos novos, discutidos nessa primeira etapa. Após a decisão sobre a melhor solução para cada caso, e sua respectiva documentação, a anotação propriamente começou. A concordância interanotadores foi medida utilizando a métrica Cohen's Kappa (Artstein, 2017). O melhor par na tarefa de anotação de relações sintáticas obteve um resultado de 95,1% de concordância, enquanto o pior par (a dupla de anotadores com mais divergências) obteve um resultado de 91,9%.

Todo o processo de revisão e avaliação foi feito por meio da ferramenta ET (de Souza & Freitas, 2021), uma estação de trabalho para busca, edição e avaliação de arquivos no formato CoNLL-U.[4] A revisão foi feita no ambiente *Interrogatório* da ET, e a avaliação foi feita no ambiente *Julgamento*.

3.1. Procedimentos de revisão

Uma vez que já é possível contar com uma anotação sintática automática de qualidade razoável para o português (veja-se os resultados apresentados em Zeman et al. (2018) para a língua portuguesa), o processo de anotação do PetroGold foi, como indicado, um processo de revisão.

A literatura sobre métodos de revisão de *corpus* é mais rica quando se trata da revisão de *treebanks*, provavelmente devido à tradição deste tipo de anotação, mas também devido à sua dificuldade. Nossa principal preocupação nesta etapa esteve na pesquisa e desenvolvimento de métodos que permitissem uma revisão capaz de encontrar erros ou inconsistências sem precisar analisar todas as palavras do *corpus*. Isto porque, se já partimos de uma anotação de qualidade média, não precisamos passar por todas as palavras do *corpus* em busca de erros, uma vez que o reconhecimento de certas formas como "artigos", "verbos" ou "advérbios" não costuma trazer dificuldades para a análise automática. Além disso, uma mesma frase pode conter erros de diferentes naturezas, o que tem como consequência dificuldade em manter o foco e a consistência, podendo tornar o processo de revisão mais suscetível a erros e mais demorado. Wallis (2003), por exemplo, recomenda trocar uma revisão linear, *token* a *token*, por uma revisão transversal, guiada pelo tipo de fenômeno linguístico, que permitiria ver os fenômenos em questão de forma ampliada e garantiria uma revisão consistente.

A revisão do PetroGold utilizou quatro estratégias para detecção de erros e inconsistências — anotações variantes, discordância entre anotações, revisão guiada por regras e revisão guiada por léxico —, descritas a seguir.

1. A estratégia de anotações variantes (ou n-gramas variantes) se baseia nos trabalhos de Dickinson & Meurers (2003a) e Dickinson & Meurers (2003b). Usada inicialmente na revisão de POS, passou a ser também aplicada na revisão de sintaxe (Boyd et al., 2008; Dickinson, 2015; de Marneffe et al., 2017). Em termos gerais, a estratégia busca inconsistências na anotação, e parte da ideia de que palavras idênticas (ou n-gramas idênticos) anotadas de maneira diferente são candidatas à inconsistência – o que nem sempre é verdade, dada a ambiguidade da língua. Levada para a anotação de dependências sintáticas, a procura por inconsistências de anotação busca detectar (i) pares de palavras idênticos que (ii) possuam uma relação (de dependência) entre eles, mas que (iii) esta relação seja diferente em cada elemento do par.

2. A estratégia de discordância entre anotações também aposta na detecção de inconsistências, mas procura inconsistências não entre sequências de palavras idênticas em um mesmo *corpus*, mas entre análises automáticas de um mesmo *corpus*, dando continuidade à estratégia aplicada em Freitas & de Souza (2023). Trata-se de uma abordagem de revisão que se inspira no procedimento humano de adjudicação das análises na anotação, quando iremos lidar com análises divergentes, e por isso a batizamos de discordância entre anotações. No entanto, substituímos análises humanas por análises automáticas, e comparamos as análises por meio de uma matriz de confusão simplificada, que nos mostra apenas divergências quanto à anotação de relações de

[4] O formato CoNLL-U é uma adaptação do formato CoNLL-X desenvolvida pelo projeto *Universal Dependencies* com o objetivo de codificar os *treebanks* que integram o projeto. Neste formato, os metadados estão indicados no início de cada sentença e, em cada sentença, os *tokens* são dispostos em sequência, um por linha. A cada *token* – em cada linha – está associada informação linguística (anotação) em 10 campos separados por tabulação, tal como lema, classe gramatical, características flexionais, relação sintática etc. Para mais informações sobre o formato, veja-se: `https://universaldependencies.org/format.html`. Acesso em 6 de nov. 2023.

dependências sintáticas. Na revisão do *corpus*, comparamos análises fornecidas por duas ferramentas capazes de produzir bons resultados — Stanza (Qi et al., 2020) e UDPipe (Straka et al., 2016) — e trabalhamos sobre a saída da ferramenta com o melhor desempenho, Stanza, chamada "anotação guia". Isto é, se na comparação entre as duas análises, a anotação guia estiver correta, não precisamos fazer nada. Se a anotação "desafiante" estiver correta, ou se nenhuma das anotações estiver correta, precisaremos efetuar a correção. A estratégia de examinar, por meio da matriz de confusão, as divergências entre análises automáticas como potenciais casos de erro traz ainda a vantagem de permitir generalizar e criar hipóteses a partir dos tipos de erros — ou inconsistências — mais comuns, facilitando a percepção de padrões de erros, o que por sua vez (i) acelera a correção (erros de um mesmo tipo tendem a ter correções parecidas); (ii) permite o desenvolvimento de regras para auxiliar a detecção e correção, e (iii) contribui para o aperfeiçoamento da documentação, caso os erros sejam decorrência de lacunas das diretrizes de anotação. Além disso, cada pessoa responsável pela revisão pode selecionar um grupo de divergências (ou de confusões) para rever, o que dá mais agilidade ao processo e menos chances para inconsistências. Por fim, esta abordagem se baseia na hipótese de que duas ferramentas não cometerão os mesmos erros, ou seja, se há convergência, é porque existe acerto, o que nem sempre se verifica.[5]

3. A revisão guiada por regras linguísticas utiliza desde as regras de validação disponibilizadas pela equipe do projeto UD[6] até, e principalmente, regras relacionadas a fenômenos mais específicos e relevantes apenas para a língua portuguesa, que criamos ao longo do processo de revisão. As regras foram desenvolvidas tendo como base (i) o conhecimento da gramática UD, (ii) o conhecimento da gramática do português, (iii) a exploração dos erros mais comuns da anotação automática, e (iv) a exploração de erros detectados pelos outros métodos, e que puderam se transformar em regras de detecção de erros. A lista de regras desenvolvidas certamente não é exaustiva e padrões atípicos nem sempre são erros na anotação, sendo necessária verificação humana para corrigir os erros identificados. O conjunto de regras inclui regras que buscam eventuais erros formais introduzidos pelos anotadores durante a revisão do *corpus*, e regras específicas da estrutura da língua portuguesa. Como exemplo do primeiro tipo, temos uma regra que busca por ciclos na árvore sintática, quando um *token* é dependente de si mesmo. É um erro grave, que inutiliza a árvore sintática da frase, mas que pode ser introduzido sem que a pessoa responsável pela anotação perceba. Como exemplo do segundo tipo, temos uma regra que sinaliza quando há diferença entre os traços morfológicos de um adjetivo e do substantivo que ele modifica. Ao longo da construção do PetroGold, foram criadas 64 regras, que podem ser aplicadas na correção de outros *treebanks* de português que sigam a abordagem UD.[7]

4. A revisão guiada por léxico utilizou o PortiLexicon-UD (Lopes et al., 2022), léxico disponibilizado pelo projeto POeTiSA, e foi aplicada na revisão de lemas, anotação de POS e de características morfológicas. O PortiLexicon-UD é um recurso público e inclui 1.221.218 entradas (palavras ambíguas foram contabilizadas como entradas diferentes quando tinham classificações distintas) com informações morfológicas de acordo com a gramática UD. No processo de revisão, comparamos todas as entradas do léxico com todos os *tokens* do PetroGold. Como esperado, nem todas as palavras do PetroGold existiam no léxico, devido sobretudo à presença de terminologias da área. Quando havia pareamento entre formas do PortiLexicon-UD e formas do PetroGold, verificamos se alguma das anotações do PetroGold divergia do léxico e, se fosse um erro, corrigimos o *corpus*. Na comparação, desconsideramos as palavras cuja POS fosse PROPN, NUM ou X (nomes próprios, numerais e palavras estrangeiras, respectivamente), pois poucas dessas palavras existiam no léxico. Esta forma de revisão foi utilizada apenas na preparação da terceira e última versão do *corpus*.

[5] Buscamos, dentre os *tokens* corrigidos do *corpus*, quantos estavam invisíveis na matriz de confusão porque correspondiam a convergências entre as anotações. O resultado desta análise indicou que as convergências correspondiam a acertos em 94,7%, 95,3%, 98,4% nos casos de identificação do elemento que governa a relação sintática (dephead), do tipo de relação sintática (deprel) e de POS, respectivamente.

[6] As regras de validação UD estão em https://github.com/UniversalDependencies/tools/blob/master/validate.py.

[7] As regras criadas para o PetroGold estão em <https://github.com/alvelvis/ACDC-UD/blob/master/validar_UD.txt>

3.2. Criação do PetroGold

A primeira fase de elaboração do *treebank* padrão ouro produziu um pequeno *corpus*, chamado Petro1. No Petro1, que tem 22.288 *tokens*/652 frases e é composto por resumos e introduções de documentos que compõem o *corpus* Petrolês, todos os *tokens* foram revistos. Em seguida, fizemos uma segunda rodada de anotação, com um conjunto de cerca de 5 mil *tokens*, chamado Petro 2, no qual também todas as frases foram revistas.

Tanto Petro1 quanto Petro2 são pequenos para treinar um modelo de dependências sintáticas, mas podem ser utilizados para avaliação. Assim, este material viabilizou a realização de testes em busca do melhor modelo de anotação sintática, fundamental no processo de revisão do PetroGold.

A segunda fase de revisão já operou sobre o *corpus* PetroGold, que foi anotado automaticamente com um modelo customizado do anotador Stanza treinado com um material que combinava o *corpus* Bosque-UD (Rademaker et al., 2017), composto por textos jornalísticos, e os *corpora* Petro1 e Petro2. O material passou por uma revisão cuidadosa e deu origem ao PetroGold v1.

A terceira fase de revisão enfatizou o tratamento de fenômenos linguísticos específicos, decorrentes sobretudo de mudanças nas diretrizes do projeto UD (PetroGold v2), e a última fase de revisão também foi pautada pela revisão de fenômenos linguísticos pontuais, dando origem à versão final do *corpus*, o PetroGold v3. A contribuição de cada método de revisão, considerando a versão final do *corpus*, está na Figura 1, e o processo de construção do PetroGold v3, bem como uma avaliação sobre a contribuição de cada método de revisão, estão detalhadamente descritos por de Souza (2023).

3.3. Avaliação

Ao longo de cada versão fomos medindo — indiretamente — a consistência interna da anotação por meio de uma avaliação intrínseca, usando sempre a mesma versão da ferramenta UDPipe. Ao longo das versões 2 e 3, foram extensivamente revistas etiquetas (estruturas linguísticas) consideradas opcionais em UD.

Por exemplo, a etiqueta *obl:arg* não é obrigatória em UD, uma vez que é uma especificação da etiqueta *obl*, destinada a elementos nominais preposicionados associados a verbos – no caso do subtipo *obl:arg*, os elementos são aqueles que nossa tradição gramatical convencionou chamar de "objetos indiretos" (como a palavra "sinergias" no exemplo (1)).

(1) *obl:arg*: Aprimoramentos nesta tecnologia **resultarão** em **sinergias** entre a tecnologia proposta e aumento de recuperação de petróleo, sendo uma recomendação para futuros desenvolvimentos.

Fizemos o mesmo para as etiquetas *expl:pv*, *expl:impers* e *expl:pass*, que são especificações da etiqueta mais geral *expl*, atribuída, entre outros casos, ao pronome expletivo *–se*. Dada a tradição gramatical do português de especificar o tipo de *–se* entre índice de indeterminação do sujeito (*expl:impers*, frase (2), pronome apassivador (*expl:pass*, frase (3) e partícula integrante do verbo pronominal (*expl:pv*, frase 4), optamos por incluir as especificações no *treebank*.

(2) *expl:impers*: A princípio, **trabalhou-se** com a hipótese de que, quanto maior o percentual de esmectita de uma argila, maior seria sua afinidade pelo metal.

(3) Somente ao misturar as duas fases é que **se adiciona** o agente modificador.

(4) *expl:pv*: Este estudo **se baseia** nas propriedades magnéticas dos minerais que **se concentram** nas rochas da crosta terrestre.

Se, por um lado, a introdução dessas etiquetas granulares foi motivada pelo tipo de informação linguística que codificam, que consideramos relevante para o processamento do conteúdo de textos, por outro lado, dificulta a comparação entre as versões do *corpus*, e por isso também medimos os resultados levando em conta o que chamamos de versões simplificadas, nas quais as etiquetas granulares são convertidas nas respectivas etiquetas mais gerais (por exemplo, *expl:pv*; *expl:impers* e *expl:pass* são convertidos em *expl*, a única etiqueta obrigatória). Os resultados estão na Tabela 1, onde se pode perceber um aumento de até 1,39 p.p. na métrica CLAS no que se refere ao desempenho do anotador automático treinado no PetroGold v3 com a simplficação das etiquetas.[8]

[8] A métrica UAS (*unlabeled attachment score*) avalia o acerto nas dependências sintáticas, sem levar em conta o tipo da relação sintática, a métrica LAS (*labeled attachment score*) avalia o acerto nas dependências e no tipo de relação, e a métrica CLAS (*content labeled attachment score*) avalia o acerto nas dependências e no tipo de relação, mas considera apenas as relações que se estabelecem entre palavras de conteúdo.

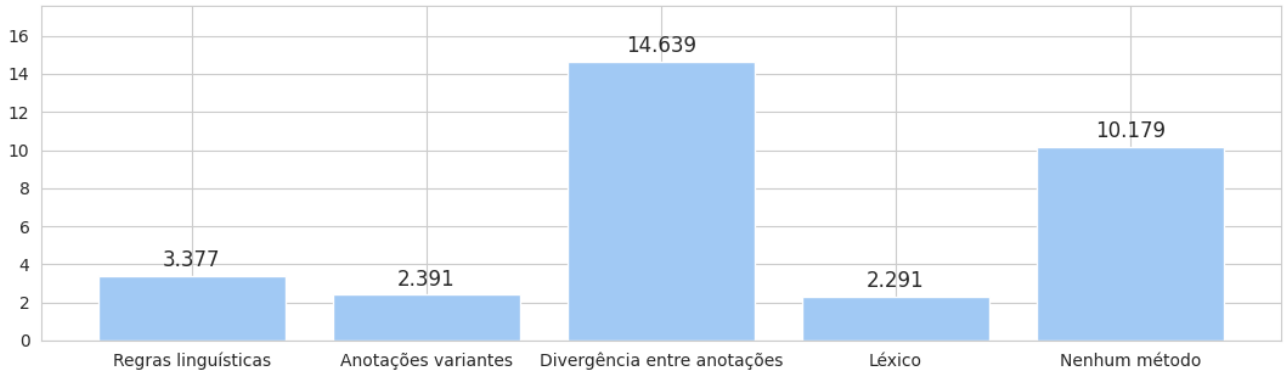

Figura 1: Distribuição da aplicação dos métodos de revisão no PetroGold v3, por *token*, em números brutos.

PetroGold	*Tokens* revistos	UPOS/simp.	UAS/simp.	LAS/simp.	CLAS/simp.
v1	—	98,19/—	90,65/—	88,53/—	82,96/——
v2	8.802	98,40/98,40	90,92/90,66	89,09/88,82	84,07/83,48
v3	9.314	98,63/98,63	91,04/92,02	89,36/90,92	84,22/85,61

Tabela 1: Evolução da anotação sintática do *corpus* PetroGold.

4. Entidades mencionadas e o PetroNer

Diferentemente do PetroGold, que contém documentos acadêmicos, o PetroNer é composto por boletins técnicos. A anotação de entidades está codificada na 9ª coluna do arquivo *CoNLL-U*,[9] e segue o formato de anotação IOB (Inside–Out–Beginning).

Frequentemente, entidades mencionadas são nomes próprios, mas como sinalizam Cohen et al. (2017) e Thompson et al. (2017), nomes próprios são menos relevantes e informativos em domínios técnicos, e é justamente a especificidade da terminologia a responsável por dificultar o processo de anotação. No PetroNer, nomes próprios continuam relevantes, mas, não são suficientes, e entidades podem ser nomes próprios, nomes comuns ou adjetivos.

No reconhecimento e classificação de entidades mencionadas, o principal desafio está na própria definição do que seja uma entidade do domínio – ou seja, o reconhecimento. Em um estudo realizado no âmbito do ACE (*Automatic Content Extraction*), uma das primeiras avaliações conjuntas sobre entidades mencionadas, uma equipe de anotadores experientes obteve concordância de apenas 82.8% na tarefa de identificação de entidades (Maynard et al., 2003). Para a língua portuguesa, no contexto do primeiro HAREM, um estudo de Mota (2007) verificou que, no que se refere à identificação (feita por pessoas), a concordância quanto às classes atribuídas ficou em torno de 45% (e 55% considerando apenas nomes próprios). Já na classificação, a concordância ficou pouco acima de 70%. Mais recentemente, na anotação do material para a tarefa 10 do SemEval (2017) (tarefa de identificação, classificação e relacionamento entre ternos-chave em publicações científicas das áreas de Ciência da Computação, Ciência de Materiais e Física), feita por estudantes de graduação e professores das referidas áreas, o índice de concordância quanto à identificação do que seriam os termos-chave do domínio ficou entre 45% e 85%, sendo a concordância igual ou maior a 60% em metade dos casos (Augenstein et al., 2017).

No PetroNer, a dificuldade na identificação foi contornada com a utilização de classes de entidades e suas instâncias definidas por um grupo de especialistas da Petrobras. Entidades são conceitos ou categorias usadas para agrupar objetos que possuem características próprias em comum. Já as instâncias são os objetos em si, que possuem as características definidas por uma entidade. Por exemplo, a entidade CAMPO corresponde, no domínio, aos campos de petróleo, que são áreas que delimitam estratos geológicos portadores de hidrocarbonetos passíveis de serem extraídos comercialmente. As instâncias *Campo de Marlim* e *Campo de Albacora* são dois exemplos de indivíduos agrupados sob o conceito de CAMPO.

[9] Originalmente esta coluna é dedicada às *Enhanced Dependencies*, mas está inutilizada no PetroNer.

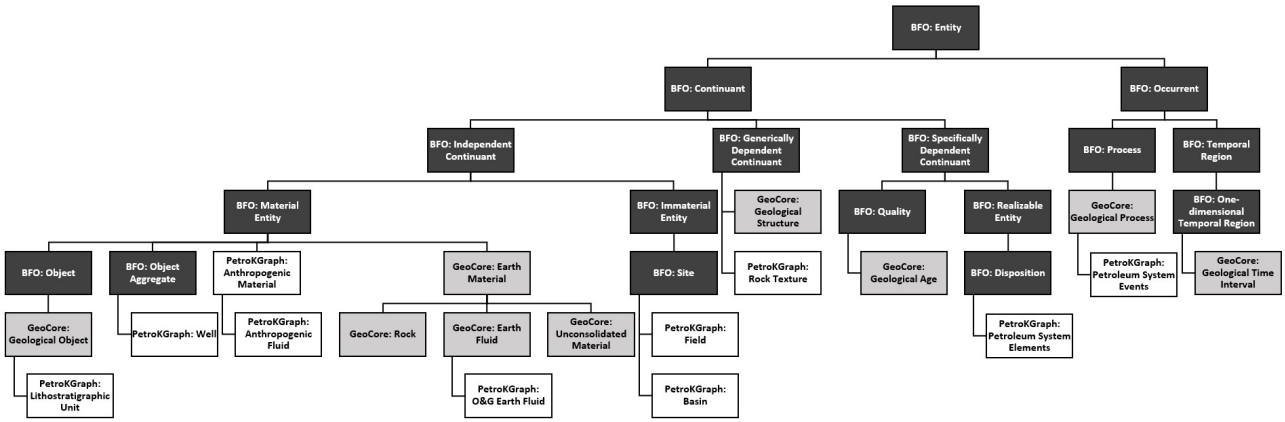

Figura 2: Árvore de relações hierárquicas do PetroKGraph.

Primeiramente, os especialistas indicaram quais tipos de entidades são importantes quando um geocientista busca por documento técnicos. Essas entidades foram então formalizadas em uma ontologia de aplicação denominada PetroKGraph, que importa e estende conceitos da *Basic Formal Ontology* (BFO)[10] e da GeoCore (Garcia et al., 2021). A BFO e a GeoCore são respectivamente ontologias de topo e de domínio (ou *core*) que possibilitam a formalização de complexos conceitos geológicos. Sempre que possível, os conceitos definidos na PetroKGraph buscaram ser compatíveis com outras ontologias desenvolvidas para o domínio de óleo e gás (Cicconeto, 2021; Cicconeto et al., 2020; Silva, 2022; Qu et al., 2023). A Figura 2 apresenta a taxonomia is_a da PetroKGraph.

Além das classes das entidades, os especialistas também compilaram listas de instâncias para cada classe. As listas não pretendiam ser exaustivas, mas abrangentes o suficiente para auxiliar no processo de anotação. Cada entidade anotada no *corpus* também recebe um identificador único referente à instância (codificado no campo *misc* do arquivo CoNLL-U, e atribuído ao primeiro *token* da entidade), além da anotação da classe da entidade, já mencionada. A Figura 3 ilustra um trecho do arquivo anotado, com destaque para a informação que codifica os identificadores de instâncias. Vale destacar que grafias alternativas (ou erros de digitação ou digitalização) de uma mesma instância recebem o mesmo identificador único, o que é valioso para tarefas de desambiguação e povoamento de grafos de conhecimento. Os trabalhos de formalização da ontologia e a construção de grafos do conhecimento oriundos dos documentos do Petrolês ainda estão em andamento e serão divulgados posteriormente.

A anotação de entidades foi feita com base em regras que utilizam informação morfossintática. Para tanto, utilizamos um modelo de anotação de dependências sintáticas gerado pela ferramenta Stanza (Qi et al., 2020) e treinado no PetroGold v2 completo.[11] Como todo o material do PetroGold foi utilizado no treinamento do modelo, para avaliar a qualidade da anotação sintática no PetroNer realizamos a revisão manual de 50 frases (1.658 *tokens*), selecionadas por terem muitos verbos (e supostamente serem mais complexas). Os resultados foram 98,37% (UPOS); 94,51% (UAS); 92,58% (LAS), e 87,23% (CLAS).[12]

O léxico inicial fornecido por profissionais de engenharia de petróleo e de geologia foi aplicado ao *corpus*, e criamos regras tanto para eliminar os casos errados como para incluir anotações que faltavam, produzindo uma anotação padrão ouro. Assim, a utilização de regras linguísticas, além de otimizar o processo de revisão, possibilitou a identificação de novas instâncias para as classes de interesse, enriquecendo o léxico inicial.

O processo de revisão do *corpus* e de criação de regras foi feito por duas linguistas (alunas dos anos finais de graduação em Letras) e contou com a supervisão direta de especialistas da área (geologia e engenharia de petróleo). O trabalho durou cerca de 8 meses e resultou na anotação de quase 20 mil entidades no PetroNer. Todo o trabalho foi auxiliado pelo ambiente *Interrogatório*, também utilizado na construção do PetroGold.

4.1. A anotação do PetroNer

A anotação automática das entidades é realizada em etapas. A primeira delas consiste na anotação do *corpus* com o léxico compilado por especialistas. O léxico contém 18 classes, com en-

[10] Basic Formal Ontology 2.0 Specification and User's Guide, https://github.com/BFO-ontology/BFO/raw/master/docs/bfo2-reference/BFO2-Reference.pdf

[11] A versão 3 ainda não estava disponível.

[12] Estas 50 frases com anotação padrão ouro estão disponibilizadas com o arquivo do PetroNer.

```
# text = Membro Mucuri, Eocretáceo da Bacia do Espirito Santo..
# sent_id = boletins-000001-7
1       Membro  Membro  PROPN   _       Gender=Masc|Number=Sing 0       root    B=UNIDADE_LITO  end_char=501|grafo=#membro_010|start_char=495
2       Mucuri  Mucuri  PROPN   _       Gender=Masc|Number=Sing 1       flat:name       I=UNIDADE_LITO  start_char=502|end_char=508
3       ,       ,       PUNCT   _       _       4       punct   0       start_char=508|end_char=509
4       Eocretáceo      Eocretáceo      PROPN   _       Gender=Masc|Number=Sing 1       conj    B=UNIDADE_CRONO|end_char=520|grafo=#LowerCretaceous|start_char=510
5-6     da      _       _       _       _       _       _       _       start_char=521|end_char=523
5       de      de      ADP     _       _       7       case    0       _
6       a       o       DET     _       Definite=Def|Gender=Fem|Number=Sing|PronType=Art        7       det     0       _
7       Bacia   Bacia   PROPN   _       Number=Sing     4       nmod    B=BACIA|end_char=529|grafo=#BASE_CD_BACIA_270|start_char=524
8-9     do      _       _       _       _       _       _       _       start_char=530|end_char=532
8       de      de      ADP     _       _       7       flat:name       I=BACIA _
9       o       o       DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       7       flat:name       I=BACIA _
10      Espirito        Espirito        PROPN   _       Number=Sing     7       flat:name       I=BACIA|start_char=533|end_char=541
11      Santo   Santo   PROPN   _       Number=Sing     7       flat:name       I=BACIA|start_char=542|end_char=547
12      .       .       PUNCT   _       _       1       punct   0       start_char=547|end_char=548
```

Figura 3: Codificação das informações no *corpus* PetroNer.

tidades do tipo BACIA e UNIDADE LITOESTRATIGRÁFICA, e 383.168 instâncias distribuídas por essas classes (nem todas foram encontradas no *corpus*). Mesmo em um domínio técnico e lidando com terminologias, a ambiguidade está presente, e por isso a fase de revisão é fundamental. Neste ponto, a tarefa de anotação pode ser vista como uma tarefa de desambiguação. A palavra *bioturbação*, por exemplo, pode pertencer à classe ESTRUTURA FÍSICA ou à classe POROSIDADE; a palavra *água* pode ser uma entidade do tipo FLUIDO DA TERRA DE INTERESSE DA INDÚSTRIA (como no caso de *água de formação*), ou ainda FLUIDO ANTROPOGÊNICO (por exemplo, em *água destilada*), ou não ser uma entidade, e apenas o contexto (ou o conhecimento especializado) será capaz de indicar a anotação correta.

Em seguida, buscamos no *corpus* — já anotado automaticamente com dependências sintáticas — a distribuição, por lemas, para cada uma das palavras anotadas com alguma das 18 classes de entidades. Este passo buscava verificar, caso a caso, o que havia sido anotado com cada etiqueta. Após a análise da lista de lemas, (i) organizamos as palavras conforme seus contextos sintáticos, criando subgrupos de revisão; (ii) identificamos as colocações associadas a cada palavra anotada, e (iii) eliminamos as etiquetas daquelas que não eram entidades. Por exemplo, as buscas pelas colocações da palavra *campo* incluíram sequências como *campo magnético*, *campo de tensões*, ou *trabalho de campo*, contextos em que a palavra "campo" não é considerada entidade.

Também criamos regras para a identificação de falsos negativos, como a busca por palavras terminadas em *-iano* ou *-oceno* que não haviam recebido a etiqueta UNIDADE CRONOESTRATIGRÁFICA, mas que deveriam, e buscas por elementos coordenados ou em relação de aposição com entidades, mas que não haviam sido anotados, como no exemplo 5, que ilustra uma entidade não anotada (*Paraná*), coordenada a uma entidade anotada. Após as análises, as devidas correções foram realizadas por meio de regras.

(5) Entre 1981 e 1990 fez parte da equipe de avaliação de perfis e teste nas *Bacias de Campos*[BACIA] e do **Paraná**, dedicando-se à área de hidrodinâmica e hidroquímica.

Classes com (i) palavras muito frequentes e polissêmicas, como *água* e *óleo*, bem como (ii) entidades do tipo propriedades, como *pioneiro*, *especial* ou *de extensão*, foram inteiramente anotadas por meio de regras — dispensando a fase inicial de aplicação do léxico. No primeiro caso, as regras tornam a anotação menos custosa, uma vez que eliminar os casos errados seria mais trabalhoso. No segundo caso, identificamos os elementos nominais modificados pelos adjetivos (ou nomes modificadores) de interesse e apenas nesses contextos os adjetivos eram anotados como entidade.

Durante todo o processo de revisão, os casos duvidosos foram resolvidos por especialistas da área. Para facilitar a análise, usamos a anotação sintática e organizamos as candidatas a entidade segundo seus perfis lexicográficos, isto é, a palavra candidata à entidade associada a seus núcleos e/ou modificadores. Este procedimento de agrupamento linguístico das palavras candidatas também acelera o processo de criação de regras de revisão, explicitando contextos em que etiquetas devem ser eliminadas, incluídas ou modificadas. A Figura 4 ilustra o processo de construção do PetroNer, e a Tabela 2 traz a totalização de *tokens* e instâncias distintas anotadas para cada entidade encontrada no *corpus*.[13]

[13] Foram também anotadas as entidades TIPO_POROSIDADE, POÇO_T, POÇO_Q e POÇO_R que representam, respectivamente, o tipo de porosidade encontrado nas rochas, o tipo do poço, a classificação do poço segundo sua finalidade (e.g., *estratigráfico* ou *pioneiro*), e o papel do poço no desenvolvimento do campo de petróleo (*produtor*). Essas entidades apresentaram quantidade pouco significativa de *tokens* anotados e, por isso, não foram formalizadas na ontologia nem utilizadas para treinamento de modelos de PLN.

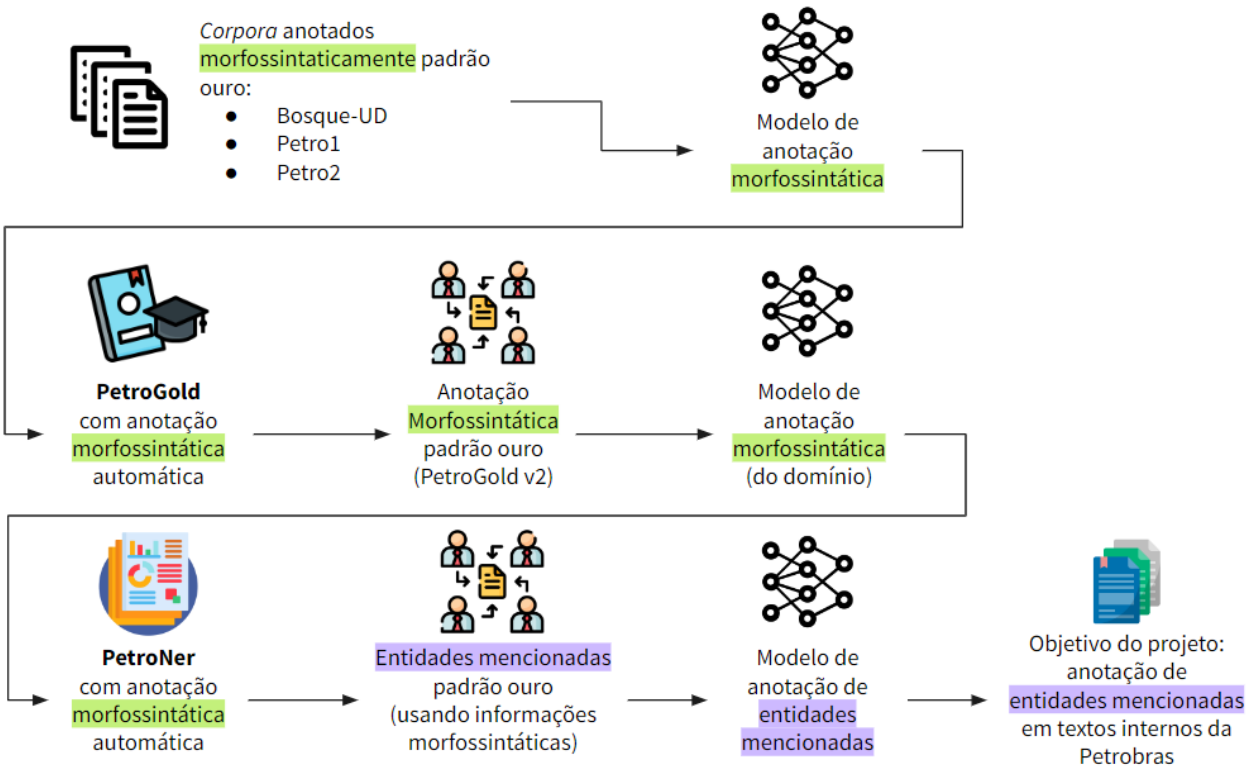

Figura 4: Fluxograma de construção do PetroNer.

4.2. Anotador baseado em regras

O processo de anotação de entidades tem início com a anotação dos arquivos CoNLL-U com as palavras vindas do léxico inicial. Nesta fase, são necessários alguns ajustes no nível de pré-processamento, como a remoção de acentos e a normalização, passando todas as palavras do léxico para minúsculas. Assim, se uma entrada do léxico é "BACIA DE CAMPOS", no texto do arquivo CoNLL-U tanto a ocorrência "bacia de Campos" quanto "BACIA DE CAMPOS" serão anotadas conforme a classe indicada no léxico, sendo a primeira palavra do tipo "B" (*begin*) e as demais do tipo "I" (*in*). Caberá às regras de correção, em outra etapa, corrigir estes casos específicos em que pode ter ocorrido algum erro. Além disso, em algumas palavras, como *Campos* e *Santos*, foi necessário manipular as lematizações produzidas pelo modelo de anotação morfossintática para que não fossem transformadas em *campo* e *santo*, respectivamente.

O léxico disponibilizado pelos especialistas de domínio é abrangente: contém um total de 382.120 entradas, distribuídas em 16 classes, e que serão aplicadas no *corpus*, como mencionamos, sem nenhuma restrição de contexto. A ideia é que a anotação via léxico seja a mais abrangente possível, abarcando todos os possíveis casos de entidades mencionadas do *corpus* — o que, inevitavelmente, irá incluir falsos positivos. É nesse contexto que se justifica a utilização de regras de anotação e de revisão linguisticamente motivadas, desenvolvidas, por um lado (e minoritariamente), para complementar os léxicos a partir das estruturas linguísticas que podem indicar novas entidades não previstas nos léxicos, e por outro lado (e majoritariamente), para realizar a revisão da anotação inicial proveniente dos léxicos, eliminando os falsos positivos. São as regras, portanto, que garantem a precisão da anotação de entidades mencionadas.

As regras de anotação e revisão foram organizadas em blocos conforme o tipo de ação que executam no processo de anotação, e para cada bloco é feita uma nova iteração em todo o *corpus*, tornando mais difícil que alguma regra deixe de ser aplicada devido à ordem em que está disposta no código. Existem 6 blocos (ou seis tipos de ação de anotação), descritos a seguir. Todas as regras foram criadas e testadas no *corpus* com a ferramenta *Interrogatório*, e só então foram incluídas em algum dos blocos.

1. Correções sintáticas — são as primeiras regras aplicadas, pois algumas das regras de anotação semântica dependem dos outros níveis linguísticos. São regras que corrigem erros da anotação morfossintática e, embora sejam as primeiras na ordem de aplicação,

Classe na ontologia	Entidade no PetroNer	*Tokens* no PetroNer	Instâncias distintas no PetroNer
PetroKGraph: Basin	BACIA	4.268	66
GeoCore: Geological Age and Geological Time Interval	UNIDADE_CRONO	3.104	134
GeoCore: Rock	ROCHA	2.772	79
GeoCore: Geological Structure	ESTRUTURA_FÍSICA	2.041	64
PetroKGraph: Lithostratigraphic Unit	UNIDADE_LITO	1.496	217
PetroKGraph: O&G Earth Fluid	FLUIDODATERRA_i	1.378	7
PetroKGraph: Well	POÇO	1.234	194
GeoCore: Unconsolidated Material	NÃOCONSOLID	1.035	11
PetroKGraph: Field	CAMPO	707	111
PetroKGraph: Petroleum System Events	EVENTO_PETRO	257	3
PetroKGraph: Earth Fluid	FLUIDODATERRA_o	246	1
PetroKGraph: Petroleum System Elements	ELEMENTO_PETRO	214	3
PetroKGraph: Athropogenic Fluid	FLUIDO	175	1
PetroKGraph: Rock Texture	TEXTURA	140	23

Tabela 2: Distribuição de entidades no PetroNer.

podem ser criadas a qualquer momento do processo de anotação, sempre que um erro de anotação do modelo de dependências sintáticas for encontrado. Este bloco de regras torna o PetroNer um material parcialmente revisto no nível morfossintático.

2. Adição de entidades novas — este bloco destina-se às entidades descobertas que não estavam no léxico inicial.

3. Expansão — regras que procuram candidatas a entidades a partir de certas estruturas linguísticas, como coordenação e aposto.

4. Revisão — regras que corrigem erros derivados da anotação inicial dos léxicos ou de outras regras anteriores.

5. Regras de mudança de entidade — regras aplicadas apenas no final e que têm como condição alguma palavra já anotada como entidade, isto é, já com alguma classificação semântica específica.

6. Regras de limpeza final (acabamentos) — regras que eliminam erros detectáveis pelo formato da anotação IOB. Por exemplo, uma regra que elimina a anotação da entidade I (*intermediate*) se o *token* anterior tem anotação de entidade de uma classe diferente (aplicada em casos como Bacia_B=BACIA de_I=CAMPO Campos_I=CAMPO, no qual a entidade do primeiro *token* é diferente das demais).

Ao final do processo, foram anotadas quase 20 mil entidades e foram "descobertas" 299 novas instâncias de entidades, enriquecendo 10 das 18 classes — por "descobertas", nos referimos às entidades que não haviam sido previstas nos léxicos especializados mas que, por meio da estrutura linguística das frases, conseguimos identificar como potenciais entidades relevantes para

o domínio, sendo posteriormente validadas pelos especialistas.

As classes que mais foram enriquecidas com dados do *corpus* foram *unidade cronoestratigráfica*, que recebeu 100 novas palavras, *unidade litoestratigráfica*, que recebeu 64, *poço* (56) e *bacia* (48).[14]

4.3. PetroNer como um *benchmark*

Dispondo de um *corpus* padrão ouro, e uma vez que a anotação foi inteiramente realizada por meio de regras, investigamos o comportamento das regras no *corpus* e simulamos o desempenho de um anotador baseado em regras.

A limitação de ferramentas de anotação baseadas exclusivamente em regras linguísticas quando aplicadas a textos inéditos é a impossibilidade de prever exatamente o que vai acontecer em uma boa parcela dos dados. Para anotar cerca de 20 mil entidades foram criadas quase 2 mil regras, e mais da metade delas (56.2%) foi aplicada apenas uma vez.

A tabela 3 traz a distribuição da frequência de aplicação das regras menos frequentes. A análise dos casos permite algumas observações sobre a abrangência das regras: (1) regras que foram aplicadas apenas uma vez (56% delas) são aquelas que realizam correções locais, em frases específicas, de erros da anotação via léxico que não puderam ser transformados em regras de correção gerais; (2) 33% das regras desenvolvidas, por sua vez, foram aplicadas mais de 4 vezes no *corpus*, indicando que são regras para fenômenos que se repetem com maior frequência no PetroNer e que, portanto, são potencialmente replicáveis em outros *corpora* do mesmo tipo, seja para revisar a anotação dos léxicos especializados, seja para encontrar novas entidades a partir da estrutura linguística dos textos.

Freq. aplicação	Qtd. regras
<= 4	1504 (77%)
<= 3	1427 (73%)
<= 2	1316 (68%)
= 1	1091 (56%)

Tabela 3: Distribuição da frequência de aplicação das regras menos frequentes na anotação do *corpus* PetroNer.

A fim de simular o desempenho de um anotador baseado apenas em regras, e, portanto, com poder limitado de generalização, anotamos o PetroNer utilizando as regras aplicadas com frequência maior ou igual a três. Os resultados estão na Tabela 4, assim como para regras com frequências próximas, para comparação. É importante mencionar, entretanto, que as regras foram criadas sem a preocupação sistemática de evitar redundância; o foco estava em corrigir problemas evitando produzir novos erros. Assim, é possível que uma análise cuidadosa das regras diminuísse a quantidade de regras com frequência de aplicação 1 e 2.

Tipo de regra	Precisão	Abrangência
freq. >= 4	98,1%	97,7%
freq. >= 3	98,4%	98,2%
freq. >= 2	98,9%	98,6%

Tabela 4: Distribuição da frequência das regras menos frequentes na aotação do *corpus* PetroNer.

As mesmas regras foram aplicadas a um conjunto de documentos que não podemos tornar públicos, a fim de verificar possibilidades de generalização das regras. O conjunto tem cerca de 250 mil *tokens* e foi anotado com o mesmo modelo de anotação de dependências usado no PetroNer. No entanto, em diversos documentos a qualidade do texto (originalmente PDFs transformados em texto simples) era ruim, prejudicando todo o fluxo de anotação. Neste contexto, a anotação baseada em regras conseguiu 94,1% de precisão e 85% de abrangência. Das 1.939 regras derivadas da criação do PetroNer, 1.594 não foram aplicadas nenhuma vez (eram regras de correção de anotação direcionadas para frases específicas que só existiam no PetroNer) — e 1.123 novas regras foram criadas ao longo do processo de tornar este material um pequeno *corpus* padrão ouro para avaliação.

Os resultados mostram que, embora uma grande quantidade de regras não tenha sido aplicada neste novo material — de fato, a maioria das regras —, aquelas que foram aplicadas, porque apresentam um alto grau de generalização, foram suficientes para produzir um outro *corpus* anotado com alta precisão (94,1%) e abrangência (85%).

[14] Todo processo de aplicação das regras, bem como as regras em si, estão disponíveis em https://github.com/alvelvis/Regras-PetroNer.

5. O idioma Petrolês

Embora pertençam ao gênero técnico-científico, PetroGold[15] e PetroNer têm origens distintas: o primeiro contém teses e dissertações; o segundo, boletins e relatórios técnicos, textos mais próximos daqueles para os quais o projeto Petrolês foi criado. Na busca por uma caracterização linguística do "idioma" Petrolês, verificamos o quanto o PetroNer se aproxima linguisticamente do PetroGold, *corpus* do qual "deriva" sintaticamente, o que traria consequências para a utilização de modelos treinados no segundo e aplicados ao primeiro, e o quanto ambos se aproximam (ou distanciam) de um *corpus* jornalístico como o Bosque, tendo em vista especialidades da linguagem técnica. A Tabela 5 apresenta características dos *corpora*.

Como podemos observar, a média de orações por frase é bem próxima entre os três *corpora*. Os números mais altos do PetroGold se devem a decisões de sentenciação que fizemos no pre-processamento, unindo itemizados como uma única frase caso o separador fosse vírgula ou ponto e vírgula. O PetroNer, ainda que também contenha listas e itens, não passou por um tratamento textual tão cuidadoso, e o Bosque, por sua vez, quase não contém este tipo de estrutura devido à natureza dos textos jornalísticos. Importante mencionar que na contagem de orações foram descartados os verbos que participam de expressões multipalavras, como em "ou *seja*", "a *partir* de" ou "*visto* que". Já quanto à quantidade de frases com pelo menos uma oração, os números se distanciam, e os números destacadamente mais baixos no PetroNer se devem à manutenção, neste *corpus*, de referências bibliográficas — estrutura linguística que costuma ser escassa em verbos — listadas ao final de cada relatório ou boletim. O mesmo motivo explica a alta proporção de frases sem oração no PetroNer. No PetroGold, apesar das referências bibliográficas terem sido excluídas, os títulos de capítulos, seções e subseções – estruturas que também costumam ser escassas em verbos, e que também estão presentes no PetroNer – explicam a diferença numérica para o Bosque. Corroborando essa prevalência verbal do Bosque, está a sua frequência relativa de verbos, que é de 10,5%, superior às frequências do PetroGold (9,2%) e do PetroNer (7,0%), como indica a tabela 6, complementar à Tabela 1. Por fim, assim como a baixa ocorrência de orações, a voz passiva — elemento capaz de impessoalizar textos — aparece como um elemento que contrasta o texto técnico-científico do jornalístico, sendo típico do primeiro.

A Tabela 6 traz a distribuição das classes de palavras mais frequentes em cada *corpus*, e a Tabela 7 a distribuição das relações sintáticas mais frequentes. Como podemos observar na Tabela 6, as principais diferenças estão na frequência dos nomes próprios, mais alta no PetroNer, como esperado, e na frequência dos verbos, mais alta no Bosque, como já discutido. Na comparação entre as relações sintáticas mais frequentes, a diferença mais visível está na classe *flat:name*, usada para os nomes próprios compostos, e por isso mais alta no PetroNer. Também é interessante constatar que, por um lado, a distribuição das classes no PetroGold e no PetroNer é muito próxima, e um dos pontos que os diferencia do Bosque é a frequência mais alta, neste último, da relação *nsubj*, utilizada para anotar sujeitos de orações ativas. Como já comentamos, a impessoalização aparece como uma característica de textos técnico-científicos, o que leva a uma diminuição na frequência de sujeitos explícitos neste tipo de material. Outra diferença está na alta frequência — 6ª posição — tanto no PetroGold quanto no PetroNer, de elementos coordenados, indicados pela relação *conj*. No Bosque, *conj* ocupa a 10 posição.

De um ponto de vista lexical, a comparação entre os adjetivos usados no PetroGold e no Bosque mostrou que, em relação a um total de 3.858 adjetivos, 653 (16,9%) eram compartilhados entre ambos os *corpora*, 1.242 eram exclusivos do PetroGold, e 1963 exclusivos do Bosque. Analisando os 50 adjetivos mais frequentes dentre os 1.242 adjetivos exclusivos do PetroGold, verificamos que 62% deles correspondem a adjetivos terminológicos, como "sedimentar", "deposicional" ou "estratigráfico", sugerindo que os termos específicos do domínio compõem a maior parte dos adjetivos que não são compartilhados pelo Bosque.

Na comparação entre os 50 verbos mais frequentes no PetroNer e no PetroGold, encontramos 72% de convergência. PetroGold e Bosque, no entanto, compartilham apenas 32% dos verbos, e PetroNer e Bosque apenas 30%.

O terceiro verbo mais frequente no Bosque, "dizer", sequer aparece entre os 50 mais frequentes do material Petrolês. O mesmo para o verbo "afirmar", também típico de discurso relatado, muito presente no jornalismo, com posição 16 no Bosque, e quase inexistente no Petrolês (posições 251 no PetroGold e 350 no PetroNer). Verbos típicos do Petrolês, por sua vez, como "observar"

[15] Todas as comparações foram feitas utilizando o PetroGold v2 pois o modelo de anotação sintática do PetroNer foi treinado nele.

	PetroGold v2	**PetroNer**	**Bosque-UD v2.12**
Número de orações	22.278	38.699	21.491
Frases com pelo menos uma oração	7.623 (85,2% de 8.949 frases)	14.267 (59,4% de 24.035 frases)	8.611 (92,0% de 9.357 frases)
Média de orações por frase	**2.9**	**2.7**	**2.5**
Número de frases sem oração	1.326 (14,8% de 8.949 frases)	9.768 (40,6% de 24.035 frases)	756 (8,0% de 9.357 frases)
Número de orações na voz passiva	4.245 (19% de 22.278 orações)	5.771 (14,9% de 38.699 orações)	1.681 (7,8% de 21.491 orações)

Tabela 5: Características dos *corpora*.

	PetroGold v2			**PetroNer**			**Bosque-UD**		
#	**upos**	**freq.**	**(%)**	**#**	**upos**	**freq.**	**(%)**	**#**	**upos**	**freq.**	**(%)**
1	NOUN		26,1	**1**	NOUN		21,8	**1**	NOUN		21,0
2	ADP		19,7	**2**	ADP		17,2	**2**	DET		17,7
3	DET		16,5	**3**	PROPN		16,1	**3**	ADP		17,1
4	VERB		9,2	**4**	DET		13,9	**4**	VERB		10,5
5	ADJ		7,7	**5**	ADJ		8,1	**5**	PROPN		9,5
6	PROPN		5,4	**6**	VERB		7,0	**6**	ADJ		5,8
7	NUM		3,3	**7**	NUM		4,5	**7**	ADV		4,3
8	AUX		3,0	**8**	CCONJ		2,7	**8**	PRON		3,8
9	CCONJ		2,9	**9**	ADV		2,6	**9**	SCONJ		2,7
10	ADV		2,8	**10**	PRON		2,2	**10**	CCONJ		2,7
11	PRON		2,6	**11**	AUX		1,8	**11**	AUX		2,5
12	SCONJ		0,8	**12**	SCONJ		0,8	**12**	NUM		2,4

Tabela 6: Distribuição das classes de palavras mais frequentes nos corpora.

(posição 4 no PetroGold e 5 no PetroNer) e "ocorrer" (posição 7 no PetroGold, e 3 no PetroNer), ocupam as posições 196 e 76 do Bosque.

Considerando os 50 substantivos comuns mais frequentes, há somente 54% de convergências entre PetroNer e PetroGold. Esta queda é resultado da constituição dos *corpora*, e se explica, novamente, pela presença de referências bibliográficas no PetroNer, mas não no PetroGold. Quando analisamos apenas as 50 palavras mais frequentes classificadas como nomes próprios, a diferença aumenta, e temos apenas 26% de convergência. Além das referências bibliográficas (que incluem nomes próprios de pessoas e de locais), a própria natureza dos boletins e relatórios, com mais entidades da área, explica a diferença.

Apesar da diferença nas classes nominais, PetroGold e PetroNer têm muitas semelhanças, por um lado, e divergências com um *corpus* jornalístico, por outro. A semelhança contribui para a boa performance do modelo de anotação morfossintática aplicado no PetroNer, e as diferenças entre petrolês e texto jornalístico ajudam a entender diferenças de desempenho entre analisadores automáticos preparados para lidar com um ou outro tipo de texto.

6. Considerações finais

Apresentamos aqui alguns recursos para o PLN de língua portuguesa, desenvolvidos ao longo do projeto Petrolês e sumarizados na Tabela 8.

Além de criarem condições para identificação e classificação de entidades de um domínio, recursos como o PetroNer e PetroGold permitem avançar com pesquisas na área, ajudando a responder questões como (i) se e quanto a incorporação de *embeddings* do domínio, como PetroVec (Gomes et al., 2021), facilita a tarefa de identificação de entidades; (ii) se e quanto a incorporação de dependências sintáticas facilita a tarefa de identificação de entidades; (iii) se e quanto a incorporação de *embeddings* do domínio facilita a tarefa de dependências sintáticas; (iv) se e quanto modelos gerais de língua têm o desempenho piorado quando aplicados a textos de um domínio específico.

#	PetroGold v2 deprel	freq. (%)	#	PetroNer deprel	freq. (%)	#	Bosque-UD deprel	freq. (%)
1	case	17,7	**1**	case	15,5	**1**	det	17,5
2	det	16,1	**2**	det	13,4	**2**	case	16,6
3	nmod	11,4	**3**	flat:name	11,2	**3**	nmod	9,5
4	amod	6,6	**4**	nmod	10,4	**4**	nsubj	5,5
5	obl	5,4	**5**	amod	7,1	**5**	obl	5,1
6	conj	4,1	**6**	conj	6,6	**6**	obj	5,0
7	root	4,0	**7**	root	4,8	**7**	amod	4,8
8	flat:name	3,4	**8**	obl	4,3	**8**	root	4,7
9	nsubj	3,3	**9**	nsubj	2,7	**9**	advmod	4,0
10	cc	2,9	**10**	cc	2,7	**10**	conj	3,3
11	obj	2,9	**11**	obj	2,4	**11**	flat:name	2,9
12	advmod	2,5	**12**	advmod	2,4	**12**	cc	2,7
13	nummod	2,3	**13**	nummod	2,3	**13**	mark	2,7
14	acl	2,1	**14**	appos	2,3	**14**	xcomp	1,6
15	aux:pass	1,6	**15**	acl	1,8	**15**	appos	1,6

Tabela 7: Distribuição das relações sintáticas mais frequentes nos corpora.

Corpus	***Tokens***	**Frases**	**Anotação**	**Anotação padrão ouro**
PetroTok	38.472	1.139	não se aplica	tokenização e sentenciação
Petro1	22.288	652	lema, pos, morf, sintaxe	lema, pos, morf, sintaxe
Petro2	5.248	166	lema, pos, morf, sintaxe	lema, pos, morf, sintaxe
PetroGold	250.605	8.946	lema, pos, morf, sintaxe	lema, pos, morf, sintaxe
PetroNer	615.418	24.035	lema, pos, morf, sintaxe, entidades	entidades

Tabela 8: Características dos *corpora* do projeto Petrolês.

Durante o desenvolvimento dos recursos, investimos na dimensão metodológica da criação de recursos, e medimos *modos de fazer*. A construção do PetroGold permitiu investigar maneiras de buscar erros de anotação no *corpus* (e construir *treebanks* de maneira eficiente) e criou uma série de regras para detecção de erros em *treebanks* de língua portuguesa que sigam o formato e a gramática UD; a construção do PetroNer permitiu um estudo inicial sobre aplicação de regras, e propôs medidas que podem funcionar como *baseline* da tarefa de anotação de entidades.

Com a onipresença de LLMs (*Large Language Models*), pode parecer antiquado o trabalho de preparação de recursos como esses. No entanto, quando aplicados a áreas de especialidade cujos conteúdos não estão facilmente acessíveis, os modelos gerais tendem a ter um fraco desempenho, que por sua vez pode ser ajustado/customizado desde que existam os recursos adequados. Mas o próprio desempenho dos modelos só pode ser avaliado se existem meios para isso.

Agradecimentos

Esse trabalho foi realizado com o apoio da Petrobras e da Agência Nacional do Petróleo e Gás Natural e Biocombustíveis (ANP).

Referências

Amaral, Daniela, Sandra Collovini, Anny Figueira, Renata Vieira, Renata Vieira & Marco Gonzalez. 2017. Processo de construção de um corpus anotado com entidades geológicas visando. Em *11th Brazilian Symposium in Information and Human Language Technology*, 63–72.

Artstein, Ron. 2017. Inter-annotator agreement. Em *Handbook of linguistic annotation*, 297–313. Springer. 10.1007/978-94-024-0881-2_11.

Augenstein, Isabelle, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman & Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE – extracting keyphrases and relations from scientific publications. Em *11th International*

Workshop on Semantic Evaluation (SemEval-2017), 546–555. 10.18653/v1/S17-2091.

Boyd, Adriane, Markus Dickinson & W Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation* 6(2). 113–137. 10.1007/s11168-008-9051-9.

Cavalcanti, Tatiana, Aline Silveira, Elvis de Souza & Cláudia Freitas. 2021. Os limites da palavra e da sentença no processamento automático de textos. *Revista Brasileira de Iniciação Científica* 8. e021033.

Cicconeto, Fernando. 2021. *GeoReservoir: An ontology for deep-marine depositional system description*: UFRGS. Tese de Mestrado.

Cicconeto, Fernando, Lucas Valadares Vieira, Mara Abel, Renata dos Santos Alvarenga & Joel Luis Carbonera. 2020. A spatial relation ontology for deep-water depositional system description in geology. Em *XIII Seminar on Ontology Research in Brazil and IV Doctoral and Masters Consortium on Ontologies (ONTOBRAS)*, 35–47.

Cohen, Kevin Bretonnel, Karin M. Verspoor, Karën Fort, Christopher S. Funk, Michael Bada, Martha Palmer & Lawrence E. Hunter. 2017. The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain. Em *Handbook of Linguistic Annotation*, 1379–1394. Springer. 10.1007/978-94-024-0881-2_53.

Cordeiro, Fábio Corrêa. 2020. Petrolês-como construir um corpus especializado em óleo e gás em português. PUC-Rio. Monografia para obtenção do título de Especialização.

Dickinson, Markus. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass* 9(3). 119–138. 10.1111/lnc3.12129.

Dickinson, Markus & Detmar Meurers. 2003a. Detecting errors in part-of-speech annotation. Em *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 107–114.

Dickinson, Markus & W Detmar Meurers. 2003b. Detecting inconsistencies in treebanks. *IEEE Transactions on Learning Technologies* 3. 45–56.

Freitas, Cláudia & Elvis de Souza. 2023. A study on methods for revising dependency treebanks: in search of gold. *Language Resources and Evaluation* 1–21. 10.1007/s10579-023-09653-4.

Gábor, Kata, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna & Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. Em *12th International Workshop on Semantic Evaluation*, 679–688. 10.18653/v1/S18-1111.

Garcia, Luan Fonseca, Mara Abel, Michel Perrin & Renata dos Santos Alvarenga. 2021. The GeoCore ontology: A core ontology for general use in geology. *Computers & Geosciences* 135. 104387. 10.1016/j.cageo.2019.104387.

Gildea, Daniel & Daniel Jurafsky. 2000. Automatic labeling of semantic roles. Em *38th Annual Meeting on Association for Computational Linguistics (ACL)*, 512–520. 10.3115/1075218.1075283.

Gomes, Diogo da Silva Magalhães, Fábio Corrêa Cordeiro, Bernardo Scapini Consoli, Nikolas Lacerda Santos, Viviane Pereira Moreira, Renata Vieira, Silvia Moraes & Alexandre Gonçalves Evsukoff. 2021. Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry* 124. 103347. 10.1016/j.compind.2020.103347.

Kim, J.D., T. Ohta, Y. Tateisi & J. Tsujii. 2003. GENIA corpus–semantically annotated corpus for bio-textmining. *Bioinformatics* 19(1). i182–i182. 10.1093/bioinformatics/btg1023.

Lewkowycz, Aitor, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari & Vedant Misra. 2022. Solving quantitative reasoning problems with language models. ArXiv [cs.CL]. 10.48550/arXiv.2206.14858.

Lopes, Lucelene, Magali Sanches Duran, Paulo Fernandes & Thiago Pardo. 2022. PortiLexicon-UD: a portuguese lexical resource according to universal dependencies model. Em *13th Language Resources and Evaluation Conference (LREC)*, 6635–6643.

Lopes, Lucelene & Renata Vieira. 2013. Building domain specific parsed corpora in Portuguese language. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 1–12.

de Marneffe, Marie-Catherine, Matias Grioni, Jenna Kanerva & Filip Ginter. 2017. Assessing the annotation consistency of the Univer-

sal Dependencies corpora. Em *4th International Conference on Dependency Linguistics (DepLing)*, 108–115.

de Marneffe, Marie-Catherine, Christopher D Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational linguistics* 47(2). 255–308. 10.1162/coli_a_00402.

Maynard, Diana, Kalina Bontcheva & Hamish Cunningham. 2003. Towards a semantic extraction of named entities. *Recent Advances in Natural Language Processing (RANLP)* 257–263.

Mota, Cristina. 2007. Estudo preliminar para a avaliação de REM em Português. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 19–34. Linguateca.

Mota, Cristina & Diana Santos. 2009. Corte e costura no AC/DC: auxiliando a melhoria da anotação nos corpos. Relatório técnico. Linguateca. http://hdl.handle.net/10400.26/20540.

Nooralahzadeh, Farhad, Lilja Øvrelid & Jan Tore Lønning. 2018. SIRIUS-LTG-UiO at SemEval-2018 task 7: Convolutional neural networks with shortest dependency paths for semantic relation extraction and classification in scientific papers. Em *12th International Workshop on Semantic Evaluation*, 805–810. 10.18653/v1/S18-1128.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. Em *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 101–108. 10.18653/v1/2020.acl-demos.14.

Qu, Yuanwei, Michel Perrin, Anita Torabi, Mara Abel & Martin Giese. 2023. GeoFault: A well-founded fault ontology for interoperability in geological modeling. *Computers & Geosciences* 105478. 10.1016/j.cageo.2023.105478.

Rademaker, Alexandre, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick & Valeria De Paiva. 2017. Universal dependencies for Portuguese. Em *4th International Conference on Dependency Linguistics (DepLing)*, 197–206.

Samuel, David, Andrey Kutuzov, Lilja Øvrelid & Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. Em *Findings of the Association for Computational Linguistics (EACL)*, 1954–1974. 10.18653/v1/2023.findings-eacl.146.

Santos, Diana. 2011. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language* 3(2). 113–128.

Santos, Diana & Cristina Mota. 2010. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. Em *7th International Conference on Language Resources and Evaluation (LREC)*, 1437–1444.

Santos, Diana & Luís Sarmento. 2003. O projecto AC/DC: acesso a corpora/disponibilização de corpora. *XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL)* 705–717.

Silva, Patricia Ferreira da. 2022. *ResRiskOnto: an application ontology for risks in the petroleum reservoir domain*: PUC-Rio. Tese de Mestrado.

de Souza, Elvis. 2023. *Construção e avaliação de um treebank padrão ouro*: PUC-Rio. Tese de Mestrado. 10.17771/PUCRio.acad.62693.

de Souza, Elvis & Cláudia Freitas. 2021. ET: A workstation for querying, editing and evaluating annotated corpora. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 35–41. 10.18653/v1/2021.emnlp-demo.5.

de Souza, Elvis & Cláudia Freitas. 2023. Explorando variações no tagset e na anotação universal dependencies (UD) para português: Possibilidades e resultados com base no treebank petrogold. Em *XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 125–134.

de Souza, Elvis, Aline Silveira, Tatiana Cavalcanti, Maria Castro & Cláudia Freitas. 2021. Petrogold – corpus padrão ouro para o domínio do petróleo. Em *XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 29–38. 10.5753/stil.2021.17781.

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28.

Straka, Milan, Jan Hajic & Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization,

morphological analysis, pos tagging and parsing. Em *10th International Conference on Language Resources and Evaluation (LREC)*, 4290–4297.

Thompson, Paul, Sophia Ananiadou & Jun'ichi Tsujii. 2017. The GENIA corpus: Annotation levels and applications. Em *Handbook of Linguistic Annotation*, 1395–1432. Springer. 10.1007/978-94-024-0881-2_54.

Wallis, Sean. 2003. Completing parsed corpora. Em *Treebanks: Building and Using Parsed Corpora*, 61–71. Springer Netherlands. 10.1007/978-94-010-0201-1_4.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre & Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. Em *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. 10.18653/v1/K18-2001.

Proposta recebida em Março 2023 e aceite para publicação em Julho 2023.

Uma revisão para o Reconhecimento de Entidades Nomeadas aplicado à língua portuguesa

A survey of Named Entity Recogntion applied to Portuguese

Andressa Vieira e Silva
Universidade de São Paulo

Resumo

O Reconhecimento de Entidades Nomeadas (REN) é a tarefa de identificação e classificação automática de entidades em um texto, tais como nomes de pessoas, lugares e organizações. Essa é uma tarefa importante em Processamento de Língua Natural, servindo como base de diversas aplicações, como tradução automática e sistemas de pergunta-e-resposta. Desde seu surgimento na década de 90, a tarefa passou por diversos fases com relação à abordagem computacional, indo dos sistemas baseados em regras manuais aos modelos de redes neurais.

Este artigo traz uma revisão da tarefa de REN considerando aplicações em textos de língua portuguesa. Apresenta-se um panorama geral da tarefa, traçando um histórico das principais iniciativas para promovê-la, dos recursos linguísticos e computacionais disponíveis e das abordagens já avaliadas para REN para o português. Por fim, apresenta-se uma discussão do cenário geral em que a tarefa se encontra e as considerações finais de análise.

Palavras chave

reconhecimento de entidades nomeadas, português

Abstract

Named Entity Recognition is the task of identifying and classifying Named Entities in a text, such as names of people, places and organizations. This is an important task in Natural Language Processing (NLP), serving as the basis for several tasks, such as automatic translation and question answering systems. Since its emergence in the 1990s, the task has gone through several periods in relation to the computational approach, ranging from rule-based systems to the neural network models.

This article presents a review of the NER task considering applications in Portuguese language texts. It presents an overview of the task, tracing a history of the main initiatives to promote the task, the linguistic and computational resources available and the approaches already applied to NER for Portuguese. Finally, a discussion of the general scenario in which the task is and the final analysis considerations is given.

Keywords

named entity recognition, Portuguese

1. Introdução

O Reconhecimento de Entidades Nomeadas[1] (REN) é uma tarefa voltada para a identificação e classificação de termos referentes a entidades em um texto, como nomes de lugares, pessoas, instâncias temporais etc. A tarefa surgiu na década de 90 como um dos tópicos de investigação da 6ª *Message Understanding Conference* (MUC-6) (Grishman & Sundheim, 1996), cujo objetivo era recuperar informações relevantes a partir de dados não-estruturados, como textos jornalísticos.

Desse modo, o REN surgiu como um dos ramos de Extração de Informação (EI) focado para a coleta das entidades de um texto. Diferente de termos com função puramente sintática, como preposições e artigos, as Entidades Nomeadas (EN) fornecem muitas pistas a respeito do conteúdo de um texto, podendo ser usadas para identificar os personagens em um livro, os nomes de países citados em uma notícia, o autor e ano de publicação de um texto etc.

Por essas razões, o reconhecimento de entidades se tornou uma etapa essencial em diversas tarefas de Processamento de Língua Natural (PLN), como tradução automática (Babych & Hartley, 2003; Li et al., 2020b), pergunta-e-resposta (Toral et al., 2005; Mollá et al., 2006) e resolução de correferências entre sintagmas (Dai et al., 2019; Gao et al., 2020). Para citar um exemplo, em tradução automática não é comum que os nomes de pessoas e países sejam traduzidos, então é importante identificá-los para que o modelo possa processá-los corretamente.

Outra tarefa associada ao REN é a Extração de Relações entre Entidades, do inglês *Entity Re-*

[1] Em inglês "Named Entity Recognition" (NER).

DOI: 10.21814/lm.15.2.396

lation Extraction, que consiste em identificar correlações entre entidades em um texto, como entre pessoa-e-organização e organização-e-lugar (Bach & Badaskar, 2007). Um exemplo, em "A sede da Apple fica na Califórnia", há uma relação de organização-e-lugar entre "sede da Apple" e "Califórnia". Essa não é uma tarefa simples, pois o seu desempenho depende da classificação correta das entidades do texto. Para o português, ela foi investigada no Segundo HAREM (Mota & Santos, 2008) e no IberLEF 2019 (Collovini et al., 2019).

O REN passou por diversas mudanças de paradigma no quesito de abordagem computacional, indo dos modelos baseados em regras ou léxicos aos modelos de aprendizado estatístico e, por fim, às rede neurais profundas. Ao longo desse período, produziu-se uma considerável literatura de revisão sobre tarefa. Um exemplo é o célebre artigo de Nadeau & Sekine (2007), em que são apresentadas discussões sobre a definição da tarefa, além de técnicas e algoritmos de aprendizado comuns na época.

Na última década, observou-se um aumento no número de artigos de revisão, principalmente aqueles voltados para ramos específicos do REN, como métodos de aprendizado de profundo (Yadav & Bethard, 2019; Li et al., 2020a) e aplicações na área de Biomedicina (Campos et al., 2012; Alshaikhdeeb & Ahmad, 2016), muitas dessas dedicadas a sistemas e recursos voltados para a língua inglesa.

Para o português, existe uma literatura de trabalhos investigativos sobre o REN conduzidos pelos pesquisadores da Linguateca[2] (Santos & Cardoso, 2007; Mota et al., 2007; Freitas et al., 2010; Mota & Santos, 2008), além de artigos para a comparação de algoritmos de aprendizado de máquina (Milidiú et al., 2007; Pellucci et al., 2011) e ferramentas (Amaral et al., 2014; Pires et al., 2017) aplicados ao REN. Mas trabalhos de revisão são datados e as pesquisas na área continuam avançando, portanto são necessárias novas revisões para a validação e comparação de técnicas mais recentes que surgiram, como os modelos de redes neurais profundas, e para a compreensão do que mudou e do que precisa ser trabalhado para novos avanços.

Este artigo tem por objetivo fornecer um panorama geral do desenvolvimento da tarefa de Reconhecimento de Entidades Nomeadas tendo em vista sua aplicação no processamento de dados em língua portuguesa. Serão apresentados os eventos promovidos para o REN, os recursos linguísticos e computacionais disponíveis e as pesquisas que têm sido feitas na área. Ademais, apresenta-se uma discussão a respeito dos desafios e caminhos futuros para pesquisas e, por fim, as considerações finais.

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

2. Definições da tarefa

Quando foi apresentada no MUC-6 (Grishman & Sundheim, 1996), definiu-se a tarefa como o reconhecimento de nomes de pessoas, lugares e organizações em textos, sendo assim a classificação de tipos de entidades definidos *a priori*. Essas três categorias acabaram se tornando as mais usuais entre os trabalhos em REN (Nadeau & Sekine, 2007), chamadas coletivamente de "ENAMEX". O MUC-6 também abrange categorias temporais, como data e tempo, e numéricas, como expressões monetárias e percentuais.

Na literatura em Português, a tarefa pode ser encontrada com duas nomenclaturas distintas: "Reconhecimento de Entidades Nomeadas" e "Reconhecimento de Entidades Mencionadas". O primeiro termo advém de uma tradução literal do nome adotado em inglês "Named Entity Recognition" e tem aparecido em diversos trabalhos na área (Amaral, 2017; Júnior et al., 2016; Mota et al., 2021; Pellucci et al., 2011). O segundo foi a adaptação inicialmente proposta para se referir à tarefa, em que "entidade mencionada" se refere às "entidades com nome próprio" (Santos & Cardoso, 2007).

As primeiras discussões a respeito do REN para a língua portuguesa estão ligadas ao HAREM, que será apresentado em mais detalhes na Seção 3.1. Diferente do MUC, o HAREM propõe um esquema de classificação para nomes próprios em geral, sem restrição a determinadas categorias (Santos, 2007). Isto é, parte-se da busca das entidades presentes nos textos em português, para depois definir-se o conjunto de categorias a partir dos exemplos encontrados.

O modelo de classificação do HAREM se baseia na ideia de vagueza da língua, em que um conceito não tem uma denotação fixa, mas depende do contexto para ser definido (Santos & Cardoso, 2007, p. 45). Sendo assim, não temos uma relação de um-para-um entre nome e objeto denotado. Em outras palavras, o mesmo nome pode se referir a mais de um objeto e é necessário o contexto para desambiguar sua referência. Por exemplo, temos nomes próprios que podem se referir a uma pessoa ou a um lugar cujo nome homenageia uma pessoa, como acontece com diversos nomes de ruas e prédios. Isso tem impacto no esquema de anotação do corpus, já que um mesmo nome pode ter mais de uma classificação possível pelo contexto.

Além disso, o HAREM considera os casos de metonímia, em que um nome tipicamente usado para designar determinado objeto é usado no lugar de outro com o qual mantém uma relação, como substituições de lugar-por-povo, empresa-por-produto (Santos & Cardoso, 2007, p. 47). No exemplo "O Brasil vai jogar na próxima semana na Copa do Mundo", "Brasil" não se refere ao país, mas aos jogadores da seleção brasileira, portanto seria ideal classificá-lo como "Pessoa". Isso não ocorre no MUC, em que uma entidade permanece com sua classificação prototípica, mesmo que o contexto forneça outra interpretação.[3] Desse modo, a perspectiva dada à tarefa mudou do MUC em relação ao HAREM, que expandiu o número de categorias de entidades e permitiu variações de classificação de acordo com o contexto de ocorrência.

A tarefa também mudou em outros aspectos conforme foi sendo aplicada em áreas específicas, como Medicina e Química, para a classificação nomes de substâncias, doenças, medicamentos, entre outros. Para citar um trabalho, Ferreira et al. (2010) buscam entidades relacionadas a diagnósticos médicos, classificando expressões como "diabetes controlado" e "alto nível de colesterol". Portanto, a tarefa não está mais restrita à classificação de nomes próprios, tendo uma aplicação muito mais ampla e diversa.

No trabalho de Marrero et al. (2013), os autores examinam e comparam diversas propostas de definição de Entidade Nomeada da literatura, apresentando análises baseadas em cunho gramatical, semântico e filosófico. No entanto, nenhuma das definições encontradas é boa o suficiente para delimitar o escopo da tarefa, o que os faz chegar à conclusão de que as Entidades Nomeadas serão definidas em razão do propósito de aplicação da tarefa.

Baseado nessa análise, não haveria uma definição pré-estabelecida para o que seriam "Entidades Nomeadas", mas *propostas de classificação*. Essas podem variar de acordo com a quantidade de categorias, o tipo (classes genéricas ou especializadas) e a organização (hierárquica ou não-hierárquica). Isso fica evidente quando se compara diferentes corpora: o CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang & De Meulder, 2003) considera quatro categorias de entidades (Pessoa, Local, Organização e Diversos) enquanto o HAREM (Santos & Cardoso, 2007; Mota & Santos, 2008) adota uma classificação hierárquica, com dez categorias principais divididas em subcategorias. Em abordagens de domínio aberto, como a Web, que visam classificar uma grande quantidade e diversidade de entidades, o número de classes pode ser ainda maior, como é o caso do modelo proposto por Sekine & Nobata (2004), que estabelece uma ontologia contendo cerca de 200 categorias de entidades.

[3] No MUC, o exemplo citado teria "Brasil" sendo classificado como "Lugar".

3. Iniciativas de fomento do REN

Desde seu surgimento, o REN ganhou muito espaço nas pesquisas em PLN. Entre 2000 e 2008, várias conferências importantes direcionaram mesas especificamente para trabalhos sobre a tarefa, entre elas o CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang & De Meulder, 2003) e o ACE (Doddington et al., 2004). No cenário atual, o REN é tópico na maioria dos eventos em PLN. Aqui, destacam-se duas iniciativas importantes que para a discussão e produção de recursos para o Reconhecimento de Entidades Nomeadas em língua portuguesa: o HAREM e o IberLEF 2019.

3.1. HAREM

A primeira grande iniciativa que fomentou o crescimento de pesquisas em REN aplicado ao Português foi o HAREM (Avaliação de sistemas de Reconhecimento de Entidades Mencionadas), uma parceria entre pesquisadores promovida pela equipe da Linguateca com o intuito de viabilizar encontros voltados para desenvolver e avaliar técnicas e sistemas para a classificação de nomes próprios em língua portuguesa. O HAREM teve duas edições, a primeira delas foi dividida em duas etapas: o Primeiro HAREM e o MiniHAREM (Santos & Cardoso, 2007), que ocorreram em 2005 e 2006, respectivamente; a segunda edição ocorreu em 2008, ficando conhecida como Segundo HAREM (Mota & Santos, 2008).

Ao todo, dez equipes participaram do HAREM submetendo seus sistemas para a avaliação. Essa foi feita com base nas Coleções Douradas (CD) do HAREM, um conjunto de corpora compostos de textos em português de vários países lusófonos que foram anotados manualmente para a tarefa. As CDs do HAREM possuem anotação hierárquica, com dez tipos de entidades principais (Pessoa, Local, Organização, Tempo, Valor, Obra, Acontecimento, Abstração, Coisa e Outro), cada qual com suas respectivas subcategorias. Os sistemas foram avaliados de acordo com diferentes métricas para medir o desempenho na identificação de entidades, na classificação morfológica (como gênero e número) e na classificação semântica (correspondente à categoria da entidade).

Os resultados apresentados ao fim do Primeiro (Santos & Cardoso, 2007) e Segundo HAREM (Mota & Santos, 2008) mostraram que a tarefa de classificação semântica foi mais desafiadora em comparação à de identificação, ficando pouco acima de 50% nos corpora avaliados. Para as ENAMEX (Pessoa, Lugar e Organização), aquela com pior desempenho geral foi Organização, o que também foi verificado em outras pesquisas (Amaral & Vieira, 2014; Santos et al., 2019). As categorias menos frequentes no corpus, como "Obra" e "Coisa", foram mais difíceis de classificar, talvez porque elas sejam mais ambíguas ou menos homogêneas em termos de padrões linguístico-ortográficos.

Além da organização desses encontros, os autores do HAREM publicaram uma extensa documentação a respeito do REN, oferecendo uma discussão sobre as dificuldades e soluções encontradas para a anotação de um corpus para o Reconhecimento de Entidades Nomeadas, as métricas de avaliação das ferramentas e metodologias para modelagem da tarefa. As Coleções Douradas produzidas foram disponibilizadas online com acesso livre, o que foi uma contribuição valorosa para a comunidade científica trabalhando com REN em português.

3.2. IberLEF 2019

O IberLEF (*Iberian Languages Evaluation Forum*) é uma campanha de avaliação conjunta voltada para diversas tarefas de processamento e compreensão de textos em línguas ibéricas. Em 2019, o IberLEF (Collovini et al., 2019) abordou a tarefa de Reconhecimento de Entidades Nomeadas como um dos tópicos do evento.

Os organizadores do IberLEF-2019 abriram uma chamada para equipes submeterem seus sistemas para avaliação. Os modelos submetidos para a competição foram avaliados em corpora de três domínios distintos: (I) textos gerais, como blogs e entrevistas, classificando cinco categorias de entidades (Pessoa, Local, Organização e Tempo e Valor), (II) dados clínicos de pacientes e (III) relatórios policiais, sendo (II) e (III) anotados apenas para Pessoa.

O IberLEF-2019 contou com a participação de cinco equipes para a competição na tarefa de REN. Os modelos variaram entre baseados em regras, baseados em aprendizado de máquina clássico, redes neurais e híbridos. Esses foram avaliados somente em relação à classificação, desconsiderando a identificação. Os resultados obtidos no IberLEF-2019 dependeram muito do tipo de corpus. No corpus policial, três modelos obtiveram bons resultados, acima de 80% medida-F. Já o desempenho no corpus clínico não foi tão promissor, uma vez que todos os modelos obtiveram menos que 50% de medida-F. No corpus geral, o melhor modelo alcançou 66,66% de medida-F, o que está longe de ser um resultado excelente.

Uma vez que os modelos foram avaliados em corpora de diferentes domínios, foi possível identificar quais foram os mais desafiadores. Como verificado, o corpus de dados clínicos se mostrou como o mais difícil, apesar de ter sido avaliado apenas para a categoria de entidade Pessoa. Isso reforça a dificuldade de aplicação da tarefa para determinados domínios. O IberLEF-2019 permitiu validar modelos recentes, como as redes neurais, para a aplicação de REN em corpora de diferentes domínios, mostrando que houve melhoria de desempenho para a classificação de entidades em textos de domínio geral. Entretanto, ainda há muito o que ser feito para que a tarefa possa ser considerada resolvida, principalmente em relação à classificação de entidades em domínio clínico.

4. Recursos linguísticos e computacionais

Existem diversos recursos linguísticos e computacionais disponíveis em língua portuguesa que podem ser aplicados sem necessidade de muitos ajustes por aqueles interessados em REN. Nesta revisão, foram selecionados recursos (ferramentas, corpora e léxicos) disponíveis em formato aberto.

4.1. Recursos linguísticos

Os corpora estão entre os principais recursos linguísticos para o processamento automático de línguas naturais. Aqui, corpus refere-se a um conjunto de textos digitais que pode ser processado por um computador. A tarefa de REN é tipicamente aplicada em corpora não-estruturados, como artigos de texto e postagens em redes sociais. Desse modo, há uma grande quantidade de conteúdos disponíveis para análise, seja em livros digitais, na Web, em revistas, dicionários etc. Contudo, os corpora mais valiosos são aqueles que possuem anotação linguística, classificando palavras, frases ou trechos de textos para uma tarefa específica.

Em REN, a anotação consiste em identificar e classificar todos os termos correspondentes a Entidades Nomeadas. Um esquema de anotação comum para a tarefa é o chamado BIO (Begin-Inside-Outside), proposto por Ramshaw & Marcus (1999), em que a primeira palavra de uma

entidade nomeada é marcada por "B-", as demais palavras da entidade (se houver) são anotados com "I-" e as palavras considerados não-entidades são marcadas por "O". A Figura 1 traz um exemplo desse tipo de anotação, em que "PES" é abreviação para "PESSOA".

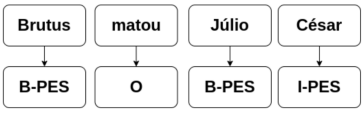

Figura 1: Representação do método de anotação BIO.

Outros esquemas de anotação para o REN são o IO (Inside-Outside), adotado no trabalho de Pirovani & Oliveira (2018), e o BILOU (Begin-Inside-Last-Outside-Unit), que aparece em Amaral & Vieira (2014). O IO distingue apenas entidades (I) de não-entidades (O), enquanto o BILOU diferencia início (B), meio (I) e fim (L) de entidades, entidades compostas de uma palavra (U) e não-entidades (O). Mas o esquema de anotação mais comum entre os trabalhos encontrados foi o BIO.

Um corpus anotado pode ser classificado em: (I) dourado, quando a sua anotação é feita manualmente e (II) prateado, quando sua anotação é feita automaticamente, sem auxílio ou revisão humana. Neste trabalho, refere-se corpora compostos de textos integral ou parcialmente escritos em Português anotados para o REN. A Tabela 1 apresenta uma comparação deles de acordo com o número de categorias de entidades, o número de entidades do corpus, o domínio e o tipo de anotação (dourado ou prateado).

Os corpora do HAREM (Santos & Cardoso, 2007; Freitas et al., 2010) são compostos principalmente de textos jornalísticos e da Web relacionados a diferentes temas, como ficção e política. O WikiNER[4] (Nothman et al., 2013), o Paramopama (Júnior et al., 2015) e o SESAME (Menezes et al., 2019) foram construídos a partir de textos de páginas da Wikipédia e anotados por meio de ferramentas computacionais. O SIGARRA (Pires, 2017) é constituído de dados da Web extraídos do sistema de informações da Universidade do Porto. As entidades são classificadas em Pessoa, Local, Organização, Data, Hora, Evento, Curso e Unidade Orgânica (por exemplo, nomes de institutos). Entre os corpora selecionados, o único composto exclusivamente de texto com linguagem da internet é o do Twitter (Peres et al., 2017), que é anotado para as categorias Pessoa, Local e Organização.

[4]O WikiNER é um corpus multilíngue, portanto apenas uma parte dele está em português.

Voltado para o domínio jurídico, o LeNER-Br (Araujo et al., 2018) é um corpus constituído de textos coletados de tribunais brasileiros e documentos legislativos. As entidades são categorizadas em Pessoa, Local, Organização, Tempo, Legislação e Jurisprudência. Já o SemClinBr (Oliveira et al., 2022) foi produzido a partir de dados clínicos de hospitais brasileiros, englobando diversas áreas de especialidade médica (cardiologia, neurologia etc.). As categorias semânticas foram baseadas no sistema UMLS,[5] que é hierárquico. Por exemplo, a categoria "Transtornos" contém subcategorias como "doença e síndrome" e "sinal ou sintoma".

Além dos corpora, a produção de léxicos, como os *gazetteers*, pode auxiliar os modelos de REN, principalmente os de abordagem híbrida ou de regras. Os *gazetteers* são repositórios contendo conjuntos de nomes próprios, por exemplo, nomes de pessoas, de doenças, de empresas etc. Eles são usados como fonte de conhecimento externo, fornecendo informações não contidas no texto que podem ser úteis para a classificação da entidade. O REPENTINO (Sarmento et al., 2006) é um léxico estruturado composto de nomes próprios extraídos do corpus WPT03.[6] Contém mais de 45.0000 exemplos de entidades divididas em 11 categorias e 97 subcategorias. Outro léxico importante é o HDBP (*Historical Dictionary of Brazilian Portuguese*), produzido por Vale et al. (2008), um dicionário de abreviações históricas.

4.2. Recursos computacionais

As ferramentas computacionais para a classificação de entidades são recursos úteis para a produção de novas ferramentas e avaliação do estado-da-arte em uma tarefa. Aqui, selecionou-se ferramentas com base em dois critérios (I) possuir modelos pré-treinados para o REN em português e (II) ser aberto. A relação de ferramentas é dada na Tabela 2.

As ferramentas encontradas estão baseadas em duas linguagens de programação muito usadas (C++ e Python). Entre essas, o spaCy é uma das mais conhecidas, projetado como uma ferramenta de aplicação industrial, conta com modelos pré-treinados em diferentes línguas, incluindo o português. O Polyglot e o Freeling também são bibliotecas com modelos pré-treinados em inúmeras línguas e tarefas. Já o

[5]https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html

[6]https://www.linguateca.pt/WPT/WPT03.html

Corpus	Categoria	Entidade	Domínio	Anotação
Primeiro HAREM	10	5.132	geral	dourado
Segundo HAREM	10	7.847	geral	dourado
Mini-HAREM	10	3.758	geral	dourado
Paramopama	4	42.769	geral	prateado
WikiNER	4	330.286	geral	prateado
SESAME	3	6.411.479	geral	prateado
SIGARRA	7	12644	geral	dourado
Twitter-NER	3	935	geral	dourado
LeNER-Br	6	44.513	Direito	dourado
SemClinBr	100	65.129	Biomedicina	dourado

Tabela 1: Corpora em língua portuguesa anotados para o REN.

Corpus	Linguagem	Referência
spaCy	Python	https://spacy.io/
NLPyPort	Python	Ferreira et al. (2019)
Freeling	C++	Carreras et al. (2004)
Polyglot	Python	Al-Rfou et al. (2015)

Tabela 2: Ferramentas para Reconhecimento de Entidades Nomeadas em português.

NLPyPort é uma biblioteca baseada no NLTK[7] que foi desenvolvida especificamente para o português. Não foram encontradas ferramentas com interface gráfica que não exijam conhecimento em programação para a utilização, mas existem alguns grupos de investigação NLX que oferecem recursos online gratuitos para o REN, como o LX-Center.[8]

A acessibilidade a modelos pré-treinados — fornecida por repositórios como o HuggingFace[9] — e a disponibilização de bibliotecas para *deep learning* — como Keras, Pytorch e Tensorflow — vêm trazendo um crescimento de interesse nas pesquisas em PLN e impulsionado a área. De acordo com Li et al. (2020a), muitos dos trabalhos reportando avanços do estado-da-arte no REN têm sido obtidos por redes neurais. Por essa razão, os recursos pré-treinados em português são essenciais, já que esses podem ser ajustados para aplicações em tarefas específicas. O HuggingFace já possui modelos de redes neurais pré-treinados em dados do português, como o BERTimbau[10] e o BioBERT.[11] O BERTimbau foi treinado no brWaC (Wagner Filho et al., 2018), um corpus sem anotação linguística composto de textos em português de diversos domínios. O BioBERT (Schneider et al., 2020) foi treinado a partir de dados clínicos de hospitais brasileiros de diversas áreas médicas, como cardiologia, neurologia e endocrinologia.

O repositório de *word embeddings* pré-treinados, disponibilizado pelo Núcleo Interdisciplinar de Linguística Computacional da Universidade de São Paulo (NILC-USP)[12] também representa um recurso valioso para pesquisas com modelos de redes neurais em português. No repositório encontram-se disponíveis modelos pré-treinados Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), Wang2Vec (Ling et al., 2015) e FastText (Bojanowski et al., 2017).

5. Abordagens para o REN

Esta seção apresenta os trabalhos encontrados a respeito de REN aplicado à língua portuguesa. Para a seleção dos artigos utilizou-se o buscador Google Scholar,[13] pesquisando pelas palavras-chave *"reconhecimento de entidades nomeadas"*, *"named entity recognition in portuguese"* e *"portuguese named entity recognition"*. Os artigos estão divididos em quatro categorias de acordo com o método de abordagem aplicado: (I) baseados em regras, (II) modelos estatísticos tradicionais, (III) redes neurais profundas e (IV) modelos híbridos.

5.1. Modelos baseados em regras

Modelos baseados em regras são projetados com base em um conjunto de diretrizes para a tomada de decisões a partir de aspectos extraídos do texto. As regras podem ter por base informações

[7] https://www.nltk.org/
[8] http://lxcenter.di.fc.ul.pt/services/en/LXServicesNer.html
[9] https://huggingface.co/
[10] https://github.com/neuralmind-ai/portuguese-bert
[11] https://github.com/HAILab-PUCPR/BioBERTpt
[12] http://www.nilc.icmc.usp.br/embeddings
[13] https://scholar.google.com.br/

linguísticas (por exemplo, padrões sintáticos e semânticos), frequência de termos, similaridade com palavras em um dicionário de referência, entre outros.

É comum dividir o REN em duas etapas nos modelos de regras: (I) identificação e (II) classificação. Na primeira, as entidades nomeadas do texto são identificadas e separadas dos termos que não constituem entidades; na segunda, as entidades identificadas são classificadas de acordo com as categorias pré-definidas pelo modelo. Essa divisão é feita porque as regras para identificação e classificação são específicas de cada etapa.

As abordagens baseadas em regras foram muito exploradas nos primeiros sistemas para Reconhecimento de Entidades Nomeadas. Em razão do Primeiro HAREM, Bick (2006) propõe o PALAVRAS-NER, um modelo de gramática contextual que aplica uma série de regras baseadas em padrões sintático-semânticos e regras de desambiguação contextual. O PALAVRAS-NER foi o sistema que obteve melhor desempenho na avaliação do Primeiro HAREM, com 58,29% de medida-F na classificação de entidades. O SIEMES (Sarmento, 2006) também participou do Primeiro HAREM, alcançando o segundo lugar com 53,30% de medida-F. Sua abordagem é baseada na similaridade de possíveis entidades presentes em um texto com aquelas listadas no *gazetteer* REPENTINO. O algoritmo faz uma busca por correspondências completas ou parciais e desambigua entre possíveis classificações a partir de um conjunto de regras contextuais.

Outro sistema de regras é o PAMPO (Rocha et al., 2016), que utiliza informações morfossintáticas das palavras para a aplicação de regras de identificação de entidades. O PAMPO aplica conhecimento externo extraído de um conjunto de listas de termos contendo, por exemplo, palavras comuns que ocorrem na fronteira de entidades. Esses termos podem estar associados a categoria da entidade, como "senhor" e "doutor", que coocorrem com nomes de pessoas. O PAMPO chegou a 73,36% de medida-F para a identificação de nomes de organizações, lugares e pessoas no corpus do HAREM.

Por sua vez, Ferreira et al. (2010) propõem um modelo baseado em ontologias e regras linguísticas para detectar entidades em cartas médicas, buscando por informações como a condição, tratamento e evolução do quadro do paciente. Os autores testaram o modelo no MedAlert, um corpus próprio composto de 90 relatórios médicos de um hospital de Portugal, reportando uma precisão de 95,0% na classificação de entidades. Outros sistemas de regras são o Rembrandt (Cardoso, 2008), o REMMA (Ferreira et al., 2008) e o CAGE (Martins et al., 2007). Os dois primeiros são guiados por páginas da Wikipédia para identificar e extrair informações para a classificação de uma EN e o último é voltado para a identificação de entidades geográficas.

Como apontado por Li et al. (2020a), um dos pontos fracos dos modelos baseados em regras é que eles tendem a obter baixa cobertura[14] por serem projetados com base em um conjunto de regras restrito e específico. Isso faz com que eles tenham baixa portabilidade, ou seja, é difícil adaptá-los a outros domínios de aplicação.

5.2. Modelos estatísticos tradicionais

Os sistemas baseados em aprendizado de máquina estatístico tratam o REN como uma tarefa de classificação multi-classes em que o objetivo é classificar cada palavra com uma etiqueta correspondendo a uma categoria de entidade ou não-entidade. Esses modelos não costumam ter módulos distintos para identificação e classificação, processando ambas etapas em um único módulo de classificação.

Assim como em modelos de regras, os algoritmos de aprendizado de máquina precisam extrair características do texto para classificá-lo, isto é, padrões relevantes que forneçam pistas de classificação para o modelo. Existem diferentes tipos de características (traços) que podem ser analisadas. Na revisão apresentada por Nadeau & Sekine (2007), os autores definem três tipos de traços: no nível da palavra (por exemplo, pistas ortográficas), baseados em listas (como os *gazzetteers*) e no nível do documento (por exemplo, contagem de palavras no texto).

Existe mais de tipo de aprendizado de máquina, mas o mais adotado em REN é o supervisionado, em que o treinamento é realizado a partir de um corpus anotado. Alguns dos algorítimos usuais são: *Hidden Markov Model* (HMM), *Decision Tree* (DT), *Support Vector Machine* (SVM) e *Conditional Random Fields* (CRF).

O NERP-CRF (Amaral & Vieira, 2014) é um modelo CRF treinado com quinze traços, divididos em ortográficos, morfossintáticos e de contexto. A Coleção Dourada do HAREM serviu para treinamento e avaliação do modelo, que mostrou alta precisão (80,77% de medida-F), mas baixa cobertura (34,59%). As autoras verificam

[14] A cobertura se refere ao número de entidades capturadas pelo modelo.

que muitos dos erros do sistema foram na delimitação de fronteiras de uma EN, o que foi causado principalmente pela preposição "de", muito recorrente em nomes próprios em português.

Júnior et al. (2015) treinam o Starford-NER,[15] uma ferramenta pré-treinada baseada em um CRF, para a classificação de entidades. São selecionados traços ortográficos e morfossemânticos. Os autores também usam *gazetteers* para nomes de pessoas, lugares e organizações. O modelo foi treinado em diferentes corpora para comparação de desempenho, tendo obtido 82,34% de medida-F com o Paramopama combinado com o HAREM.

Já o trabalho de Solorio (2007) apresenta um sistema SVM treinado para o português a partir do conhecimento de um modelo desenvolvido para o espanhol. A autora considera traços internos (como informação ortográfica) e externos, obtidos a partir de um classificador de entidades para o espanhol. Esse classificador atribui a etiqueta morfossintática e uma pré-classificação da palavra seguindo o esquema de anotação BIO. Solorio (2007) mostrou que a combinação dos traços internos e externos retornou melhores resultados do que os traços internos sozinhos, indicando que é vantajoso aproveitar de conhecimento externo advindo de modelos treinados para línguas semelhantes.

Lopes et al. (2019) treinam um algoritmo CRF para identificar nomes de doenças, sintomas, genes, entre outros, com base em um corpus próprio de artigos científicos da revista portuguesa Sinapse.[16] O sistema desenvolvido considera traços contextuais com uma janela de cinco palavras, cujas informações ortográficas e morfossintáticas são extraídas. A classificação média do modelo ficou em 72,86% de medida-F. Também utilizando um CRF, de Souza et al. (2019) classificam entidades nomeadas a partir registros de saúde de hospitais anotados manualmente. Os autores adotam o esquema de classificação UMLS, mas optam por aglomerar determinadas classes devido a um desbalanceamento do corpus, englobando as categorias em três grandes grupos (Doenças, Procedimentos e Medicamentos). Treinou-se o CRF com traços ortográficos e morfossintáticos, chegando a uma média de 55,66% de medida-F.

[15] https://nlp.stanford.edu/software/CRF-NER.shtml

[16] https://www.sinapse.pt/

5.3. Modelos de redes neurais profundas

As redes neurais profundas têm ganhado muito espaço em inúmeras aplicações em PLN. Um dos motivos para o sucesso desses modelos são as representações vetoriais *word embeddings* (Bengio et al., 2003). Os *word embeddings* forneceram uma forma eficiente para a representação de texto em modelos computacionais que só processam informações numéricas, como é o caso das redes neurais. Eles têm sido aplicados não somente em palavras, mas também em caracteres (*character embeddings*) (Santos & Guimarães, 2015; Fernandes et al., 2018). Além disso, as representações vetoriais podem representar outros tipos de traços, como ortográficos e lexicais.

As arquiteturas de redes neurais são diversas, mas as mais adotadas em PLN são as redes neurais recorrentes, como *Long Short-Term Memory* (LSTM) (Hochreiter & Schmidhuber, 1997), e, mais recentemente, as redes *Transformer* (Vaswani et al., 2017), como BERT (Devlin et al., 2018) e RoBERTa (Liu et al., 2019).

Um dos primeiros trabalhos a testar uma rede neural para o REN em português foi Santos & Guimarães (2015). Os autores propõem uma rede neural que combina *word embeddings* e *character embeddings*, chamando o modelo de CharWNN. A rede foi treinada e testada em duas línguas: português e espanhol. Em suas discussões, os autores mostram que a combinação dos dois *embeddings* (de palavra e de caractere) é mais eficiente do que essas representações usadas isoladamente. A CharWNN alcançou 65,41% de medida-F no HAREM para a classificação em todas as categorias do corpus.

Por sua vez, Castro et al. (2018) aplicam uma rede neural LSTM bidirecional (BiLSTM) com camada de saída CRF para o REN em português. Foram comparados quatro tipos de *word embeddings* (FastText, GloVe, Wang2Vec e Word2Vec), tendo o Wang2Vec obtido o melhor desempenho nos experimentos testados. Também utilizando uma rede neural BiLSTM, Fernandes et al. (2018) avaliam diferentes tipos de classificadores para fazer o processamento na saída da rede neural. Os melhores resultados foram obtidos com uma arquitetura BiLSTM seguida de uma rede neural convolucional e uma classificador de saída CRF. Assim como Castro et al. (2018), os autores obtiverem melhores resultados com vetores Wang2Vec.

Santos et al. (2019) avaliam a eficiência de combinação de *word embeddings* não-contextuais e contextuais, utilizando *Flair embeddings* (Akbik et al., 2019). Esses *embeddings* codificam in-

formações a partir das palavras vizinhas e de seus caracteres, sendo portanto uma representação de palavras e de caracteres ao mesmo tempo. A rede neural implementada foi uma BiLSTM com camada de saída CRF e o corpus de avaliação foi o HAREM.

O BERTimbau foi testado para o REN em português por Souza et al. (2019), treinado no corpus HAREM para a classificação de entidades nomeadas. Os autores comparam o resultado do BERTimbau ao do BERT treinado em um corpus multilíngue e mostram que o primeiro se sai melhor.

Peres et al. (2017) testam variações de redes BLSTM para a classificação de entidades em *tweets* no Twitter-NER. O modelo apresentado por eles se saiu melhor do que o *baseline* avaliado, um modelo CRF, mas mesmo assim alcançou pouco mais de 50% de medida-F. Entre as classes de entidades avaliadas (Pessoa, Local e Organização), "Local" foi aquela com menor desempenho.

Recentemente, houve um crescimento na aplicação de REN para domínios específicos, o que se popularizou ainda mais com as redes neurais. No domínio jurídico, Araujo et al. (2018) treinam uma rede neural BiLSTM-CRF no corpus LeNER-Br utilizando *word embeddings* GloVe. Os resultados reportados foram excelentes, ficando com uma média acima de 90% de medida-F. Por sua vez, Mota et al. (2021) testam três redes neurais: duas redes recorrentes com combinações distintas de *word embeddings* e uma rede convolucional. Os modelos foram avaliados em relação a duas categorias (Legislação e Jurisprudência), tendo como corpus de treino e teste o LeNER-Br mais um conjunto de petições iniciais de processos. O modelo com melhor desempenho foi uma rede neural recorrente com *Flair embeddings*, que alcançou média de 76% de medida-F para a classificação de entidades.

Outra área que tem chamado a atenção dos pesquisadores é a Biomedicina. Lidando com a área de neurologia, Lopes et al. (2020) apresentam duas arquiteturas BiLSTM-CRF alimentadas com *word embeddings* pré-treinados em dois domínios distintos: na Wikipédia e em dados clínicos. O modelo com os *embeddings* de domínio clínico obtiveram 75,08% de medida-F, em comparação a 74,58% obtido pelo modelo treinado em domínio geral. Esse resultado corrobora que existem diferenças nas características de entidades de domínio específico que não podem ser desconsideradas.

Schneider et al. (2020) estavam interessados em avaliar a influência do treinamento de modelos em domínio específico para o REN. Para isso, comparam o desempenho do BioBERTpt ao do BERTimbau e ao do BERT treinado em um corpus multilíngue. O BioBERT obteve 60,4% de medida-F, superando os demais modelos em mais de 2%. Aponta-se novamente que as entidades nomeadas em domínio específico têm suas particularidades que não são capturadas por modelos treinados em textos genéricos. Em um trabalho posterior, de Souza et al. (2020) aplicam o BioBERT para a classificação multinomial no corpus SemClinBr, em que uma única entidade pode ter mais de uma etiqueta em determinados contextos. Comparado aos outros modelos, o BioBERT obtém o melhor desempenho na tarefa, alcançando 3,5% a mais em relação ao *baseline* (CRF).

5.4. Modelos híbridos

Os modelos híbridos são aqueles que combinam duas ou mais abordagens distintas, como aprendizado de máquina e regras manuais. Jiang et al. (2016) aponta que os modelos híbridos são boas escolhas quando se quer combinar alta precisão e cobertura e não há disponibilidade de grande quantidade de dados anotados para o treinamento.

As abordagens híbridas podem ser das mais variadas. O trabalho de Milidiú et al. (2007) compara algumas abordagens para o REN aplicados para o português. Eles selecionam três algoritmos (HMM, SVM e *Transformation based learning*), além de um modelo *baseline* baseado em *gazetteers* e regras. Os autores realizam testes combinando os modelos entre si, mostrando que eles se saem melhor combinados do que isolados.

Ferreira et al. (2007) propõem um sistema de REN que combina dois módulos: um de regras manuais baseadas em expressões regulares para a classificação de entidades numéricas e um de aprendizado de máquina para os nomes próprios. Além disso, o modelo conta com um conjunto de léxicos de nomes para ajudar na correção de possíveis erros de classificação. Outro modelo híbrido é o CRF-LG (Pirovani & Oliveira, 2018), um sistema que combina um CRF com um conjunto de regras manuais verificadas a partir de uma gramática local usada para pré-atribuir a classificação de entidades. O CRF-LG foi avaliado no HAREM com medida-F de 57,8%.

Utilizando o método de aprendizado profundo, Júnior et al. (2016) concatenam uma rede neural LSTM com uma rede neural convolucional (CNN). A CNN gera representações de caracteres das palavras, que são combinadas a *word embeddings*. A rede LSTM é alimentada pelo vetor final gerado. Os corpora usados para treino e

teste foram o HAREM, o WikiNER e o Paramopama. Na avaliação do HAREM, o modelo obtém medida-F de 71,35% para a classificação de cinco classes de entidades (Pessoa, Local, Organização, Tempo e Valor).

Em domínio especializado, Dias et al. (2020) aplicam a tarefa de REN a dados sensíveis, capturando informações como nomes próprios, telefone, endereço e profissão de indivíduos. Para isso, os autores propõem um modelo composto de inúmeros módulos: um baseado em regras, um baseado em léxico e um de modelos estatísticos. O modelo de regras classifica entidades como números telefônicos e e-mails a partir de expressões regulares, enquanto os modelos estatísticos focam na classificação de entidades mais complexas, como pessoa, local e organização. Após a avaliação de diversos algoritmos, os autores utilizam uma rede neural BiLSTM no módulo estatístico. O modelo foi avaliado no DataSense NER Corpus, um corpus próprio anotado para a pesquisa, obtendo 83,0% de medida-F na classificação de entidades.

5.5. Panorama geral da tarefa

A Seção 5 apresentou uma revisão de trabalhos a respeito de REN para a língua portuguesa, que foram divididos por tipo de técnica adotada na abordagem. Buscou-se distinguir os trabalhos de domínio geral e específico e o corpus a que o modelo foi aplicado, já que isso reflete muito no desempenho. Para uma visão geral dos resultados, as Tabelas 3 e 4 apresentam uma comparação dos trabalhos discutidos em relação ao método utilizado, corpus de teste e medida-F obtida. As abreviações "MR", "AM", "RN" e "MB" significam "Modelo de regras", "Aprendizado de máquina", "Rede neural" e "Modelo híbrido", respectivamente.

Analisando os trabalhos da Tabela 3, vemos que a maioria foi testado no HAREM, o que o torna o principal corpus de textos de domínio geral para a avaliação de desempenho em REN. O HAREM além de ser anotado manualmente, contém muitas categorias de entidade e subcategorias, que podem ser selecionadas de acordo com o objetivo de pesquisa. Comparando os trabalhos em relação ao método de abordagem, observa-se que os modelos de redes neurais são os que vem alcançando melhores resultados. Para a avaliação com dez categorias de entidades, o modelo com o estado-da-arte é o BERTimbau (Souza et al., 2019), com 78,6% de medida-F. Já o desempenho no HAREM para apenas quatro categorias (Pessoa, Local, Organização e Tempo) foi de 80,7%, reportado por Júnior et al. (2015).

Trabalho	Método	Corpus teste	F1
Bick	MR	HAREM	58,2
Sarmento	MR	HAREM	53,3
Rocha et al.	MR	HAREM	73,6*
Martins et al.	MR	HAREM	34,1
Cardoso	MR	HAREM	56,7
Ferreira et al.	MR	HAREM	45,2
Amaral & Vieira	AM	HAREM	57,9
Júnior et al.	AM	HAREM	80,7**
Solorio	AM	Lácio Web	–
Santos & Guimarães	RN	HAREM	65,4
Castro et al.	RN	HAREM	70,3
Fernandes et al.	RN	HAREM	67,5
Santos et al.	RN	HAREM	74,6
Souza et al.	RN	HAREM	78,6
Peres et al.	RN	NER-Twitter	52,7
Milidiú et al.	MB	próprio	88,1
Ferreira et al.	MB	HAREM	92,5*
Pirovani & Oliveira	MB	HAREM	57,8
Júnior et al.	MB	HAREM	71,3**
Dias et al.	MB	próprio	83,0

* indica que o modelo foi avaliado apenas para identificação.
** indica que o modelo avaliado apenas para um subconjunto de entidades do corpus.

Tabela 3: Comparação de resultados dos trabalhos em domínio geral.

O único trabalho encontrado para a classificação de *tweets* foi o de Peres et al. (2017), que obteve 52,7% para classificação de nomes de pessoas, lugares e organizações. Esse resultado reforça a dificuldade da aplicar o REN a textos que usam linguagem da internet, como é o caso das redes sociais. Nesse tipo de meio, as entidades muitas vezes são pessoas da roda social do usuário (como amigos e família) e os nomes próprios não costumam ser escritos com letra inicial maiúscula, além de haver muitos apelidos, que são difíceis de detectar.

No domínio específico (Tabela 4), os resultados são mais difíceis de comparar, pois os trabalhos foram testados em corpora e domínios variados. Entre os corpora, o SemClinBR parece ser um dos mais desafiadores, sendo aquele com maior número de categorias e subcategorias classificadas, que podem ser confundidas entre si pelos modelos. De acordo com Schneider et al. (2020), as categorias mais difíceis foram aquelas com maior granularidade, especificidade e com vocabulários variáveis entre os hospitais, como a subcategoria "Laboratório". Na área de Direito, Araujo et al. (2018) alcançou um ótimo resultado no LeNER, com mais de 90% de medida-F.

Trabalho	Área	Método	Corpus teste	F1
Ferreira et al.	Médica	MR	MedAlert	–
Lopes et al.	Médica	AM	próprio	72,8
de Souza et al.	Médica	AM	próprio	55,6
Araujo et al.	Direito	RN	LeNER	92,5
Mota et al.	Direito	RN	próprio	76,0
Lopes et al.	Médica	RN	próprio	74,9
Schneider et al.	Médica	RN	SemClinBR	60,4
de Souza et al.	Médica	RN	SemClinBR	56,1

Tabela 4: Comparação de resultados dos trabalhos em domínio específico.

Vale ressaltar que no LeNER apenas duas categorias das seis são específicas do Direito (Jurisprudência e Legislação). Um dos pontos que favorecem na classificação de entidades em Direito é a padronização, isto é, os termos, abreviações e expressões seguem, em geral, uma fórmula, tendo textos com um estilo de escrita muito repetitivos que pode ajudar os modelos a capturarem as características das entidades.

6. Desafios e caminhos futuros da área

O Reconhecimento de Entidades Nomeadas tem alcançado bons resultados de avaliação em corpora de língua inglesa, com o estado-da-arte em 94,6% de medida-F no corpus ConLL-2003, reportado pelo trabalho de Wang et al. (2020). Já para o português, o estado-da-arte no MiniHAREM, um dos corpus mais utilizados para teste na tarefa, é de 78,6% de medida-F, obtido por Souza et al. (2019) com o BERTimbau. Isso mostra que ainda existe muito trabalho a ser feito na área. Aqui serão discutidos alguns desafios no REN em português e possíveis caminhos a serem explorados.

6.1. Criação de novos recursos

Os corpora anotados são recursos valiosos para o treinamento de algoritmos de aprendizado supervisionado. Quando se trata de redes neurais profundas, a quantidade de textos anotados precisa ser ainda maior para que o algoritmo obtenha melhores resultados. Para fins de comparação, o ConLL-2003 tem cerca de 23.499 entidades no corpus de treinamento, enquanto o Primeiro HAREM, usado geralmente para treino dos modelos, tem cerca de 5.132. Desse modo, é esperado que os modelos treinados no ConLL obtenham melhores resultados.

A criação de novos corpora anotados para o português é um caminho para avançar as pesquisas em REN. Existem diversos métodos de anotação automática através de ferramentas pré-treinadas ou de ontologias, como a DBpedia. Nesse sentido, seria interessante explorar tais recursos para a anotação, principalmente de textos em rede sociais. Como mostrado na Tabela 1, somente um dos corpora encontrados é composto completamente de textos da Web.

Li et al. (2020a) ressaltam que a classificação de entidades nomeadas em textos de redes sociais tende a ser mais desafiadora, com resultados pouco acima de 40% de medida-F. Sendo assim, é preciso produzir mais materiais para o treinamento e a avaliação em textos dessa natureza. Somente assim será possível obter modelos mais robustos, capazes de lidar com a diversidade estilística e aplicáveis no domínio da Web. Nesse sentido, a API do Twitter[17] pode ser avaliada como uma ferramenta de anotação automática de *tweets*, já que ela fornece alguns recursos prontos para a identificação de pessoas, lugares e produtos citados no texto.

Os corpora de domínios específicos também são importantes para o treinamento de modelos especializados. O domínio comercial tem recebido muita atenção em PLN nos últimos anos, principalmente para a análise de comentários de usuários a respeito de produtos. Modelos especializados em classificação de produtos já foram explorados na literatura estrangeira (Zhao & Liu, 2008; Luo et al., 2011), mas não foram encontrados trabalhos com textos em português, sendo esse um campo rico para a produção de corpora e ferramentas para REN.

6.2. Reutilização de recursos

Como a anotação manual de corpora de qualidade é um trabalho árduo e demorado, uma alternativa possível seria combinar diversos corpora existentes para o treinamento dos modelos, como foi feito no IberLef-2019. Isso pode aju-

[17] https://developer.twitter.com/en/docs/twitter-api

dar a aumentar o número de exemplos para treinamento e fornecer mais diversidade linguística e estilística, ainda mais se os textos forem de domínios e gêneros distintos. A maior dificuldade nessa estratégia é a unificação dos corpora, já que eles podem ter esquemas de anotação distintos e inconsistências que prejudiquem a classificação.

Além da quantidade de exemplos, o balanceamento do corpus é extremamente importante para que o modelo não fique enviesado pelas categorias com mais exemplos. Por isso, retirar ou concatenar categorias que não são representativas dentro do corpus também pode ajudar a melhorar o desempenho dos modelos. Por exemplo, o HAREM define 10 categorias de entidades, mas algumas delas são pouco representativas em relação às demais, o que prejudica muito o desempenho dos modelos quando avaliados em todas elas.

Por fim, a reutilização de modelos pré-treinados é um caminho promissor para avançar com as pesquisas em REN. As técnicas de *transfer learning* vêm sendo amplamente exploradas em PLN, principalmente em aplicações com redes neurais profundas (Malte & Ratadiya, 2019; Alyafeai et al., 2020). Essas técnicas exploram o conhecimento adquirido do treinamento de um modelo em um domínio para aplicá-lo em outro. Alyafeai et al. (2020) apontam que o uso de técnicas de *transfer learning* em redes neurais profundas pode diminuir a complexidade de treinamento, seja pela quantidade de parâmetros ou tempo necessário para o treinamento.

Diversos trabalhos discutidos nesta revisão fazem uso de recursos computacionais pré-treinados para a classificação de entidades nomeadas (Souza et al., 2019; Schneider et al., 2020; Fernandes et al., 2018; Castro et al., 2018; Santos et al., 2019). Mais de uma delas mostrou que modelos pré-treinados na mesma língua que a tarefa alvo apresentam melhores resultados do aqueles pré-treinados em outras línguas (Souza et al., 2019; Schneider et al., 2020), indicando que o conhecimento para a resolução do REN é muito associado à língua alvo.

Outra vantagem da reutilização de modelos pré-treinados é que eles podem ser adaptados a novos domínios de aplicação. Existem diversas maneiras de reutilizá-los, seja pelo treinamento em um corpus anotado através de *fine-tuning*, seja utilizando-os como *word embeddings* para um novo modelo ou combinados com outras técnicas para gerar modelos híbridos.

7. Considerações finais

Este artigo apresentou uma trajetória da tarefa de Reconhecimento de Entidades Nomeadas para língua portuguesa, oferecendo um panorama geral para aqueles interessados em saber mais sobre a tarefa, recursos e técnicas disponíveis. Os trabalhos foram apresentados de forma analógica ao desenvolvimento da área ao longo dos anos, partindo dos modelos de regras aos de aprendizado de máquina profundo. As pesquisas em PLN com língua portuguesa têm se dedicado ao REN há mais de 10 anos. Nesse sentido, as iniciativas de eventos sobre REN foram importantes para estabelecer a área, principalmente considerando o HAREM, que forneceu recursos computacionais e linguísticos importantes para o ponta-pé inicial da área.

No cenário atual, vê-se uma nova onda de impulsionamento nas pesquisas de REN em português, direcionada fortemente aos modelos de aprendizado profundo. Atribui-se isso à recente disponibilização de corpora anotados maiores e mais diversos, modelos computacionais pré-treinados e bibliotecas de PLN para treinamento de modelos de *deep learning*. Contudo, ainda há um longo caminho a ser percorrido até que essa seja dada como uma tarefa resolvida ou, ao menos, tenha alcançado resultados bons o suficiente para aplicação em mundo real. Apresentou-se alguns caminhos possíveis a serem adotados em via de se obter sistemas de REN mais robustos e adaptáveis para o português.

Mais do que a criação e avaliação de recursos linguísticos e computacionais, será necessário voltar um olhar mais crítico ao que já foi produzido para tentar compreender melhor as principais dificuldades da tarefa e delinear possíveis soluções. Como discutido nos trabalhos de Santos et al. (2019), Júnior et al. (2015) e Amaral & Vieira (2014), a confusão entre classes foi um erro frequente cometido pelo reconhecedor de entidades, como classificar em "pessoa" um lugar que tem seu nome homenageado a alguém, tal qual "Raposo Tavares" se referindo à rodovia. Esse é tipo de ambiguidade de classificação é um dos tipos de erros mais recorrentes entre os modelos de REN. Desse modo, ressalta-se o trabalho colaborativo entre pesquisadores para investigar tais questões, a fim de propor soluções e avaliar técnicas para superar esses problemas.

Referências

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter & Roland Vollgraf. 2019. FLAIR: an easy-to-use framework for state-of-the-art NLP. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 54–59. 10.18653/v1/N19-4010.

Al-Rfou, Rami, Vivek Kulkarni, Bryan Perozzi & Steven Skiena. 2015. POLYGLOT-NER: Massive multilingual named entity recognition. Em *SIAM International Conference on Data Mining*, 586–594. 10.1137/1.9781611974010.

Alshaikhdeeb, Basel & Kamsuriah Ahmad. 2016. Biomedical named entity recognition: a review. *International Journal on Advanced Science, Engineering and Information Technology* 6(6). 889–895. 10.18517/ijaseit.6.6.1367.

Alyafeai, Zaid, Maged Saeed AlShaibani & Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. ArXiv [cs.CL]. 10.48550/arXiv.2007.04239.

Amaral, Daniela. 2017. *Reconhecimento de entidades nomeadas na área da geologia: bacias sedimentares brasileiras*: Pontifícia Universidade Católica do Rio Grande do Sul. Tese de Doutoramento.

Amaral, Daniela, Evandro Brasil 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.

Amaral, Daniela & Renata Vieira. 2014. NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática* 6(1). 41–49.

Araujo, Pedro, Teófilo Campos, Renato Oliveira, Matheus Stauffer, Samuel Couto & Paulo Bermejo. 2018. LeNER-Br: a dataset for named entity recognition in brazilian legal text. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 313–323.

Babych, Bogdan & Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. Em *7th International EAMT workshop on MT and other language technology tools*, 1–8.

Bach, Nguyen & Sameer Badaskar. 2007. A review of relation extraction. Unpublished University Work: Literature review for the "Language and Statistics II" class.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent & Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3. 1137–1155.

Bick, Eckhard. 2006. Functional aspects in Portuguese NER. Em *International Workshop on Computational Processing of the Portuguese Language (PROPOR)*, 80–89.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)* 5. 135–146. 10.1162/tacl_a_00051.

Campos, David, Sérgio Matos & José Luís Oliveira. 2012. Biomedical named entity recognition: a survey of machine-learning tools. *Theory and Applications for Advanced Text Mining* 11. 175–195. 10.5772/51066.

Cardoso, Nuno. 2008. REMBRANDT: reconhecimento de entidades mencionadas baseado em relaçoes e análise detalhada do texto. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 195–211. Linguateca.

Carreras, Xavier, Isaac Chao, Lluís Padró & Muntsa Padró. 2004. FreeLing: An open-source suite of language analyzers. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 239–242.

Castro, Pedro, Nádia Silva & Anderson Soares. 2018. Portuguese named entity recognition using LSTM-CRF. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 83–92.

Collovini, Sandra, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro & Rafael Glauber. 2019. IberLEF 2019 Portuguese named entity recognition and relation extraction tasks. Em *Iberian Languages Evaluation Forum (IberLEF)*, 390–410.

Dai, Zeyu, Hongliang Fei & Ping Li. 2019. Coreference aware representation learning for neural named entity recognition. Em *28th International Joint Conference on Artificial Intelligence (IJCAI)*, 4946–4953. 10.24963/ijcai.2019/687.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv [cs.CL]. 10.48550/arXiv.1810.04805.

Dias, Mariana, João Boné, João C Ferreira, Ricardo Ribeiro & Rui Maia. 2020. Named entity recognition for sensitive data discovery in Portuguese. *Applied Sciences* 10(7). 2303. 10.3390/app10072303.

Doddington, George R, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel & Ralph M Weischedel. 2004. The Automatic Content Extraction (ACE) program-tasks, data, and evaluation. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 837–840.

Fernandes, Ivo, Henrique Lopes Cardoso & Eugenio Oliveira. 2018. Applying deep neural networks to named entity recognition in portuguese texts. Em *5th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 284–289. 10.1109/SNAMS.2018.8554782.

Ferreira, Eduardo, João Balsa & António Branco. 2007. Combining rule-based and statistical methods for named entity recognition in Portuguese. Em *5th Workshop em Tecnologias da Informaçao e da Linguagem Humana*, 1615–1624.

Ferreira, João, Hugo Gonçalo Oliveira & Ricardo Rodrigues. 2019. Improving NLTK for processing Portuguese. Em *8th Symposium on Languages, Applications and Technologies (SLATE)*, 18:1–18:9. 10.4230/OASIcs.SLATE.2019.18.

Ferreira, Liliana, António Teixeira & Joao Paulo Silva Cunha. 2008. REMMA: Reconhecimento de entidades mencionadas do MedAlert. Em *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 213–229. Linguateca.

Ferreira, Liliana, António Teixeira & Joao Paulo da Silva Cunha. 2010. Information extraction from portuguese hospital discharge letters. Em *VI Jornadas en Tecnologia del Habla/Speech and Languages Technologies for Iberian Languages (FALA)*, 39–42.

Freitas, Cláudia, Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota & Diana Santos. 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.

Gao, Cheng-sheng, Jun-fu Zhang, Wei-ping Li, Wen Zhao & Shi-kun Zhang. 2020. A joint model of named entity recognition and coreference resolution based on hybrid neural network. *Acta Electonica Sinica* 48(3). 442–448. 10.3969/j.issn.0372-2112.2020.03.004.

Grishman, Ralph & Beth M Sundheim. 1996. Message understanding conference-6: A brief history. Em *International Conference on Computational Linguistics (COLING)*, 466–471.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8). 1735–1780. 10.1162/neco.1997.9.8.1735.

Jiang, Ridong, Rafael E Banchs & Haizhou Li. 2016. Evaluating and combining name entity recognition systems. Em *6th Named Entity Workshop*, 21–27. 10.18653/v1/W16-2703.

Júnior, C Mendonça, Hendrik Macedo, Thiago Bispo, Flávio Santos, Nayara Silva & Luciano Barbosa. 2015. Paramopama: a brazilian-portuguese corpus for named entity recognition. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, SBC.

Júnior, Carlos, Luciano Barbosa, Hendrik Macedo & SE Sao Cristóvao. 2016. Uma arquitetura hıbrida LSTM-CNN para reconhecimento de entidades nomeadas em textos naturais em lıngua portuguesa. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 241–252.

Li, Jing, Aixin Sun, Jianglei Han & Chenliang Li. 2020a. A survey on deep learning for named entity recognition. Em *International Conference on Data Engineering*, vol. 34 1, 50–70. 10.1109/ICDE55515.2023.00335.

Li, Zhen, Dan Qu, Chaojie Xie, Wenlin Zhang & Yanxia Li. 2020b. Language model pre-training method in machine translation based on named entity recognition. *International Journal on Artificial Intelligence Tools* 29(07n08). 2040021. 10.1142/S0218213020400217.

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 (NAACL)*, 1299–1304. 10.3115/v1/N15-1142.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv [cs.CL]. 10.48550/arXiv.1907.11692.

Lopes, Fábio, César Teixeira & Hugo Gonçalo Oliveira. 2019. Named entity recognition in portuguese neurology text using CRF. Em *EPIA Conference on Artificial Intelligence*, 336–348. 10.1007/978-3-030-30241-2_29.

Lopes, Fábio, César Teixeira & Hugo Gonçalo Oliveira. 2020. Comparing different methods for named entity recognition in portuguese neurology text. *Journal of Medical Systems* 44(4). 1–20. 10.1007/s10916-020-1542-8.

Luo, Fang, Han Xiao & Weili Chang. 2011. Product named entity recognition using conditional random fields. Em *4th International Conference on Business Intelligence and Financial Engineering*, 86–89. 10.1109/BIFE.2011.101.

Malte, Aditya & Pratik Ratadiya. 2019. Evolution of transfer learning in natural language processing. ArXiv [cs.CL]. 10.48550/arXiv.1910.07370.

Marrero, Mónica, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato & Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces* 35(5). 482–489. 10.1016/j.csi.2012.09.004.

Martins, Bruno, Mário Silva & Marcirio Chaves. 2007. O sistema CaGE no HAREM-reconhecimento de entidades geográficas em textos em língua portuguesa. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área*, 98–112. Linguateca.

Menezes, Daniel, Ruy Milidiu & Pedro Savarese. 2019. Building a massive corpus for named entity recognition using free open data sources. Em *8th Brazilian Conference on Intelligent Systems (BRACIS)*, 6–11.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. ArXiv [cs.CL]. 10.48550/arXiv.1301.3781.

Milidiú, Ruy Luiz, Julio Cesar Duarte & Roberto Cavalcante. 2007. Machine learning algorithms for portuguese named entity recognition. *Inteligencia Artificial* 11(36). 67–75.

Mollá, Diego, Menno Van Zaanen & Daniel Smith. 2006. Named entity recognition for question answering. Em *Australasian Language Technology Workshop*, 51–58.

Mota, Caio, André Nascimento, Péricles Miranda, Rafael Ferreira Mello, Isabel Maldonado & José Coelho Filho. 2021. Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais. Em *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 130–140.

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

Mota, Cristina, Diana Santos & Elisabete Ranchhod. 2007. Avaliação de reconhecimento de entidades mencionadas: princípio de AREM. Em *Avaliação Conjunta: um novo paradigma no processamento computacional da língua portuguesa*, 161–175. IST Press.

Nadeau, David & Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30(1). 3–26. 10.1075/li.30.1.03nad.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy & James R Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194. 151–175. 10.1016/j.artint.2012.03.006.

Oliveira, Lucas, Ana Carolina Peters, Adalniza da Silva, Caroline Gebeluca, Yohan Gumiel, Lilian Cintho, Deborah Carvalho, Sadid Al Hasan & Claudia Moro. 2022. SemClinBr: a multi institutional and multi specialty semantically annotated corpus for portuguese clinical NLP tasks. *Journal of Biomedical Semantics* 13. 13. 10.1186/s13326-022-00269-1.

Pellucci, Paulo Roberto Simões, Renato Ribeiro de Paula, Walter Borges de Oliveira Silva & Ana Paula Ladeira. 2011. Utilização de técnicas de aprendizado de máquina no reconhecimento de entidades nomeadas no português. *e-xacta* 4(1). 73–81.

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. 10.3115/v1/D14-1162.

Peres, Rafael, Diego Esteves & Gaurav Maheshwari. 2017. Bidirectional LSTM with a context input window for named entity recognition in tweets. Em *The Knowledge Capture Conference*, 1–4. 10.1145/3148011.3154478.

Pires, André, José Devezas & Sérgio Nunes. 2017. Benchmarking named entity recognition tools for portuguese. Em *9th INForum: Simpósio de Informática*, 111–121.

Pires, André Ricardo Oliveira. 2017. *Named entity extraction from Portuguese web text*: Universidade do Porto. Tese de Mestrado.

Pirovani, Juliana & Elias Oliveira. 2018. Portuguese named entity recognition using conditional random fields and local grammars. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 4452–4456.

Ramshaw, Lance A & Mitchell P Marcus. 1999. Text chunking using transformation-based learning. Em *Natural language processing using very large corpora*, 157–176. Springer. 10.1007/978-94-017-2390-9_10.

Rocha, Conceição, Alípio Jorge, Roberta Sionara, Paula Brito, Carlos Pimenta & Solange Rezende. 2016. PAMPO: using pattern matching and pos-tagging for effective named entities recognition in portuguese. ArXiv [cs.IR]. 10.48550/arXiv.1612.09535.

Santos, Cícero Nogueira dos & Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. Em *5th Named Entity Workshop*, 25–33. 10.18653/v1/W15-3904.

Santos, Diana. 2007. O modelo semântico usado no primeiro HAREM. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área*, 43–57. Linguateca.

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

Santos, Joaquim, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini & Renata Vieira. 2019. Assessing the impact of contextual embeddings for portuguese named entity recognition. Em *8th Brazilian Conference on Intelligent Systems (BRACIS)*, 437–442. 10.1109/BRACIS.2019.00083.

Sarmento, Luis. 2006. SIEMÊS –a named-entity recognizer for portuguese relying on similarity rules. Em *International Workshop on Computational Processing of the Portuguese Language (PROPOR)*, 90–99. 10.1007/11751984_10.

Sarmento, Luís, Ana Sofia Pinto & Luís Cabral. 2006. Repentino – a wide-scope gazetteer for entity recognition in portuguese. Em *International Workshop on Computational Processing of the Portuguese Language (PROPOR)*, 31–40. 10.1007/11751984_4.

Schneider, Elisa Terumi Rubel, Joao Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro & Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt – a portuguese neural language model for clinical named entity recognition. Em *3rd Clinical Natural Language Processing Workshop*, 65–72. 10.18653/v1/2020.clinicalnlp-1.7.

Sekine, Satoshi & Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 1977–1980.

Solorio, Thamar. 2007. MALINCHE: a NER system for portuguese that reuses knowledge from Spanish. Em *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área*, 123–136. Linguateca.

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. ArXiv [cs.CL]. 10.48550/arXiv.1909.10649.

de Souza, João Vitor Andrioli, Yohan Bonescki Gumiel, Lucas Emanuel Silva & Claudia Maria Cabral Moro. 2019. Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups. Em *XIX Simpósio Brasileiro de Computação Aplicada à Saúde*, 318–323. 10.5753/sbcas.2019.6269.

de Souza, João Vitor Andrioli, Elisa Terumi Rubel Schneider, Josilaine Oliveira Cezar, Lucas Emanuel Silva, Yohan Bonescki Gumiel, Emerson Cabrera Paraiso, Douglas Teodoro & Claudia Maria Cabral Moro Barra. 2020. A multilabel approach to portuguese clinical named entity recognition. *Journal of Health Informatics* 12. 366–372.

Tjong Kim Sang, Erik F. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. Em

6th Conference on Natural Language Learning, (CoNLL).

Tjong Kim Sang, Erik F & Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Em *7th Conference on Natural Language Learning (CoNLL)*, 142–147.

Toral, Antonio, Elisa Noguera, Fernando Llopis & Rafael Munoz. 2005. Improving question answering using named entity recognition. Em *10th International Conference on Applications of Natural Language to Information Systems*, 181–191. 10.1007/11428817_17.

Vale, Oto, Arnaldo Candido, Marcelo Muniz, Clarissa Bengtson, Lívia Cucatto, Gladis Almeida, Abner Batista, Maria C Parreira, Maria Tereza Biderman & Sandra Aluísio. 2008. Building a large dictionary of abbreviations for named entity recognition in portuguese historical corpora. Em *International Conference on Language Resources and Evaluation (LREC)*, 47–54.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need 1. 5999–6008.

Wagner Filho, Jorge A, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: a new open resource for Brazilian Portuguese. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 4339–4344.

Wang, Xinyu, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang & Kewei Tu. 2020. Automated concatenation of embeddings for structured prediction. ArXiv [cs.LG] v1. 10.48550/arXiv.2010.05006.

Yadav, Vikas & Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. ArXiv [cs.CL]. 10.48550/arXiv.1910.11470.

Zhao, Jun & Feifan Liu. 2008. Product named entity recognition in Chinese text. *Language Resources and Evaluation* 42(2). 197–217. 10.1007/s10579-008-9066-8.

Projetos, Apresentam-se

Propuesta recibida en julio 2023 y aceptada para publicación en noviembre 2023.

Corpus lingüísticos del Instituto Caro y Cuervo (CLICC): una plataforma en línea para el almacenamiento, sistematización y consulta de corpus

Linguistic Corpus of the Caro y Cuervo Institute (CLICC): an online platform for corpus storage, systematization and consultation

Ruth Yanira Rubio López
Instituto Caro y Cuervo

Andrés Steban Luna Cortés
Instituto Caro y Cuervo

Nathalia Solano-Guzmán
Instituto Caro y Cuervo

Resumen

Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC) es una plataforma en línea para el almacenamiento, sistematización, administración y consulta de corpus, que nació con el objetivo de contar con un espacio para la salvaguarda de los archivos producto de las investigaciones del Instituto y que, actualmente, está disponible para que investigadores, comunidades o personas interesadas puedan publicar sus corpus sobre las lenguas de Colombia. CLICC es un espacio de acceso libre dirigido a público general y especializado interesado en explorar y contribuir a la documentación de la diversidad lingüística y cultural de Colombia. En este documento se describen sus características, funcionalidades, consultas y perspectivas futuras. También se explican los diversos ajustes que se han hecho para garantizar la publicación de corpus de distintos tipos, el respeto por los permisos y singularidad de cada corpus, y el aprovechamiento futuro de los archivos con fines diversos.

Palabras clave

corpus, herramientas de gestión de corpus, lingüística de corpus, diversidad lingüística y cultural

Abstract

Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC) is an online platform for corpus storage, systematization, administration, and query. CLICC was created with the objective of safeguarding the files produced by research conducted by the Institute. Now, it is available so that researchers, communities, or interested people can publish their corpora of Colombian languages. CLICC is a free access space open to the general and specialized public interested in exploring and contributing to the documentation of Colombia's linguistic and cultural diversity. This document describes its characteristics, functionalities, consultations, and future perspectives. It also highlights the various changes that have been made to guarantee the publication of different types of corpora, the permissions and singularity of each corpus, and the future use of the documents for various purposes.

Keywords

corpus, corpus management tools, corpus linguistics, linguistic and cultural diversity

1. Introducción

Hoy en día, los corpus son el insumo principal para multiplicidad de tareas y objetivos, entre los que podemos encontrar la investigación lingüística, la elaboración de diccionarios, la documentación, el procesamiento de lenguaje natural, que son posibles gracias a la divulgación, fácil acceso y reutilización de corpus. Como bien lo mencionan Torruella & Llisterri (1999, p. 29), "dado el esfuerzo económico y humano que supone la creación de un corpus, parece lógico pensar en que éste debe poder ser reutilizado por otros investigadores y para fines diferentes a los que fue concebido." De ahí la importancia de contar con plataformas que permitan la sistematización de los corpus, como también su divulgación y uso con fines diversos.

Algunas plataformas de este estilo son Sketch Engine (Kilgarriff et al., 2014), Gestor de Corpus (GECO) (Sierra et al., 2017), Linguistic Tools (Caminada et al., 2008), Susteainibility Platform for Linguistic Corpora and Resources (SPLICR) (Rehm et al., 2008), y la plataforma virtual que presenta este artículo, Corpus Lingüísiticos del Instituto Caro y Cuervo, en adelante CLICC.

DOI: 10.21814/lm.15.2.407

La plataforma CLICC[1] se creó con el objetivo inicial de contar con un espacio suficientemente flexible para la salvaguarda, divulgación y análisis de corpus recopilados en las distintas investigaciones realizadas en el Instituto Caro y Cuervo (Rubio et al., 2017). CLICC permite el almacenamiento, administración, análisis y publicación de corpus de distintos tipos, por parte de investigadores del Instituto. Próximamente, se espera que otros equipos de investigación, comunidades o cualquier persona interesada puedan cargar sus datos en la plataforma.

En este documento se presentan las razones de creación de CLICC y su estado actual: la estructura, funcionalidades, los corpus cargados hasta el momento, consultas, ajustes y mantenimiento que se ha realizado desde su desarrollo (2017). Además, se esclarecen algunas perspectivas para el mejoramiento de la plataforma.

2. ¿Por qué la creación de CLICC?

Tras varias décadas de investigación, el Instituto Caro y Cuervo (ICC) ha recopilado un amplio acervo de datos sobre las lenguas de Colombia (fotografías, grabaciones, manuscritos, entre otros). Muchos de estos materiales tienen potencial de convertirse en corpus, o directamente se han recopilado con los métodos de la lingüística de corpus. CLICC se creó, precisamente, para garantizar la salvaguarda de estos materiales, facilitar su almacenamiento y administración en un mismo espacio, y permitir la publicación y aprovechamiento de los corpus para fines diversos. Al respecto es importante tener en cuenta varios aspectos. Primero, CLICC funciona completamente en línea, almacena la información en los servidores del ICC, tiene una monitorización y mantenimiento por parte del grupo de las TIC, y respeta los permisos y manejo de la información de cada corpus. Así pues, sigue la misión del Instituto de salvaguardar el patrimonio lingüístico de Colombia, a la vez que ofrece un servicio gratuito con el que se disminuye el riesgo de pérdida de información valiosa. Segundo, el sistema permite en un mismo espacio virtual varias funciones que suelen encontrarse por separado en distintas herramientas: la creación y administración de los corpus, similar a *Sketch engine* (Kilgarriff et al., 2014); la adición de metadatos de las muestras y de sus hablantes, función propia de programas como *SayMore*;[2] hacer consultas de frecuencias, concordancias y colocaciones, como *AntConc* (Anthony, 2013) o *WordSmith* (Scott, 2008); y realizar consultas por los metadatos de cada corpus. Asimismo, para el análisis de corpus textuales se instaló parte del código de etiquetado morfosintáctico de *Freeling* (Padró & Stanilovsky, 2012) y *Treetagger* (Schmid, 1994) en el servidor de producción, y se creó un *script* para ejecutar estos archivos desde CLICC. Esto es lo que facilita la realización de consultas por palabra y colocaciones de corpus textuales u orales con transcripción. Tercero, la plataforma permite divulgar el corpus para el público general y hacer consultas generales y especializadas (en el usuario registrado) (Ver sección 3). Esto permite que corpus que se han recopilado para investigaciones específicas puedan ser consultados o utilizados para otras investigaciones o con otros fines; por ejemplo, para su uso en materiales de enseñanza de lenguas.

3. Estado actual de la plataforma CLICC

3.1. Estructura y funcionalidades

Para la estructuración y requerimientos iniciales de la plataforma, se tomaron como referentes los corpus orales del Atlas Lingüístico y Etnográfico de Colombia (ALEC), del Habla Culta de Bogotá (HCB) y del Español Hablado en Bogotá (EHB),[3] construidos a partir de los datos recopilados en investigaciones realizadas en el Instituto entre los años 50 y los 90 del siglo XX (Rubio & Bernal, 2019). La plataforma, entonces, estaba organizada para la publicación de estos corpus y estaba compuesta por la base de datos, la interfaz administrativa y la interfaz de usuario. La base de datos estaba organizada de acuerdo con los metadatos de informantes y de sesiones de las muestras, y las llaves primarias tenían tablas como encuestador, informantes, usuarios, entre otras (Rubio et al., 2017). Sin embargo, con el objetivo de que CLICC funcione para corpus de diversos tipos y con permisos de acceso diferentes, se añadió una interfaz para investigadores y una interfaz para usuario registrado. Actualmente, CLICC está compuesta por:

- una base de datos relacional desarrollada en MySQL y PHP;
- la interfaz de usuario general (IUG) en la que cualquier usuario puede conocer CLICC y consultar los corpus publicados;
- la interfaz de usuario registrado (IUR), que permite hacer consultas especializadas con ba-

[1] https://clicc.caroycuervo.gov.co

[2] https://software.sil.org/saymore/

[3] Para ver la lista de corpus creados y listos para cargue de datos en CLICC vea lo Cuadro 1.

TIPO	SIGLA	NOMBRE DEL CORPUS
Monolingües y orales sin transcripción	ASMYCU	Acervo de tradición oral afrocaucano "Manuel y Constanza USSA"
	CAELE/2 ORAL	Corpus del Español como Lengua Extranjera y Segunda - Oral
	EHB	Corpus del Español Hablado en Bogotá
	DDALN	Diplomado de documentación audiovisual de lenguas nativas
Monolingües y orales con transcripción	ALEC	Corpus del Atlas Lingüístico-Etnográfico de Colombia
	HCB	Corpus del Habla Culta de Bogotá
	EURP	Corpus de Espacios Urbanos de Restablecimiento Poblacional
	CLC	Corpus de habla leída y conversacional del español de Colombia
Monolingües y textuales	LVBC	Corpus Literatura de la Violencia Bipartidista en Colombia
	CLEC	Corpus Léxico del Español de Colombia
	CAELE/2 ESCRITO	Corpus del Español como Lengua Extranjera y Segunda-Escrito
	DHLC	Corpus de Documentos para la historia lingüística de Colombia
Bilingüe oral y textual	CLS	Corpus de la Lengua Sáliba

Cuadro 1: Corpus creados en CLICC

se en los metadatos de cada corpus y la descarga de archivos dependiendo de los permisos otorgados para cada corpus;

- la interfaz de usuario investigador (IUI) para el ingreso y administración de corpus;
- y la interfaz administrativa (IA) que permite la gestión de usuarios, corpus y accesos.

De la misma manera, CLICC cuenta con varios tipos de usuarios que tienen funcionalidades y permisos específicos, con el objetivo de garantizar un entorno colaborativo que respete las políticas de seguridad y las características de cada corpus, y que sea de uso intuitivo para cualquier tipo de usuario (experto o no). En cuanto a la base de datos, inició siendo relacional hasta la tercera forma, es decir, seguía unas guías de diseño más sencillas debido a las características similares que tenían los corpus orales de prueba. Sin embargo, la adición de nuevos corpus y metadatos que no compartían atributos comunes llevó a la necesidad de normalizarla hasta la sexta forma normal. Esto implicó identificar y eliminar todas las relaciones multivariadas en la base de datos para permitir el registro de una tabla que almacenara los atributos. Luego, la información se registró en otra tabla en la que se almacenarían las respuestas a los atributos previamente identificados y relacionados con la entidad de los nuevos corpus registrados.

3.2. Corpus creados en CLICC

Hasta el momento, en la plataforma se han creado 13 corpus de distintos tipos (Ver cuadro 1),que categorizamos siguiendo algunos parámetros de Torruella & Llisterri (1999). Actualmente, el investigador solicita la creación de su corpus y recibe capacitación por parte del Grupo de Lingüística de Corpus y Computacional (LICC) para realizar los siguientes procesos: sistematización de archivos, cargue y registro de metadatos y muestras; definición de metadatos de búsqueda general y especializada; ingreso de la información general del corpus; y, solicitud de la publicación del corpus. Los corpus creados en CLICC se encuentran en momentos diferentes de este proceso y, como se puede evidenciar, todos se centran en las lenguas de Colombia, por lo que la plataforma se consolida como un repositorio para la salvaguarda y estudio de la diversidad lingüística y cultural del país.

3.3. Ingreso y administración de corpus

En la IUI se ecuentran los menús para funciones vinculadas con el ingreso y administración de corpus. Primero, en la ventana principal de la IUI se encuentra el menú "Mis corpus" para escoger el corpus a trabajar. Al seleccionarlo, se pueden visualizar la cantidad y últimas sesiones consultadas por los usuarios registrados con acceso al corpus (Ver fig. 1). También, se encuentra la información general del corpus: la cantidad de hablantes, transcripciones, sesiones, etc., que se han registrado en el sistema. Esto permite visualizar el tamaño del corpus y también el avance en su construcción.

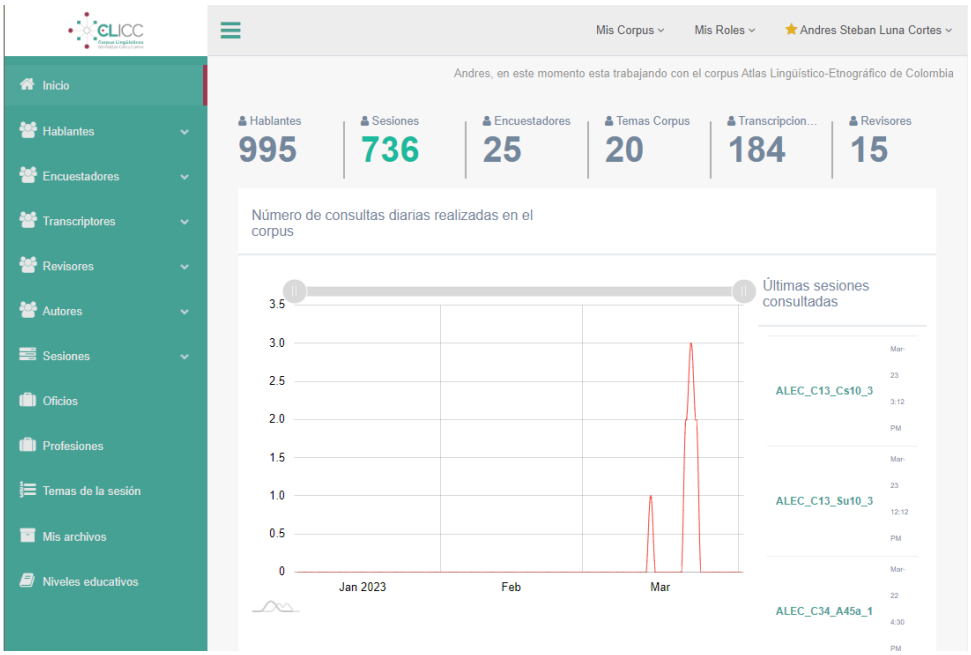

Figura 1: Interfaz de investigador CLICC

Luego, el sistema cuenta con un espacio para el almacenamiento de las muestras en distintos formatos que se encuentran en la pestaña "Mis archivos" (Ver figura 1). CLICC genera un conjunto de carpetas de acuerdo con el tipo de corpus; por ejemplo, uno oral cuenta con carpetas para los audios en formato MP3 y WAV, para las transcripciones en formato TXT, PDF, DOCx y EAF,[4] y para los anexos que incluyen una carpeta para los consentimientos informados. Los formatos tienen unos objetivos específicos dentro del sistema y no todos son obligatorios: el TXT para las consultas, los PDF que se visualizan en el usuario general, y los MP3 para agilizar la consulta de los audios, por su peso reducido. Los otros formatos pueden facilitar el aprovechamiento de las muestras de distintas maneras o con otros fines; por ejemplo, la transcripción en EAF puede facilitar el análisis lingüístico y exportación a otros formatos y el audio en WAV es más recomendado para la realización de estudios fonéticos (si el corpus es oral).

Por otro lado, en la parte izquierda de la interfaz están las distintas pestañas para el ingreso de metadatos de las muestras y de informantes (Ver figura 1). Estos metadatos se definieron a partir de los corpus iniciales de la plataforma. Sin embargo, no son obligatorios: si el corpus requiere otro tipo de metadatos se pueden definir y el sistema generará una plantilla de Excel para su registro. El ingreso de los metadatos se puede hacer diligenciando los espacios establecidos en la plataforma o a través de una plantilla de Excel, para facilitar la carga masiva de datos. Finalmente, para que las muestras, metadatos y todos los archivos de una sesión estén conectados, al terminar el ingreso de datos se realiza un proceso de registro de sesión que los vincula a todos, lo que permite las consultas del sistema. En cuanto a la administración de los corpus, al dar clic sobre el nombre de su usuario, el usuario investigador encontrará un menú que le permite:

1. gestionar los usuarios registrados que han solicitado acceso al corpus (aceptar o eliminar permisos);
2. solicitar el registro para iniciar la creación de un corpus nuevo;
3. solicitar acceso a otros corpus como usuario investigador;
4. ingresar la información de su corpus, que aparecerá en la IUG (descripción, metodología, equipo, cómo citar);
5. y seleccionar qué formatos y cuántas descargas diarias serán permitidas a los usuarios que tiene acceso al corpus (si aplica).

Teniendo en cuenta que para la creación del corpus puede haber varios usuarios involucrados y con distintas tareas, existe un usuario líder de corpus. Este será quién tenga todas las funcionalidades de administración: los numerales 1, 4 y 5 de la lista previa son exclusivos de este tipo de usuario. Este líder o líderes de corpus se marcan con una estrella amarilla al lado de su nombre (Ver Figura 1).

[4] Formato en el que quedan almacenadas las transcripciones realizadas con el programa ELAN.

Figura 2: Descripción general y consultas del corpus CorlexCO

3.4. Consultas de los corpus

La IUG es el espacio de consulta en línea para cualquier usuario interesado en conocer CLICC y consultar sus corpus. En la página web se presentan los objetivos, equipo, y guías de uso de la Plataforma. Asimismo, en la pestaña de cada corpus el usuario encontrará la descripción general del corpus, la sección de consultas, la metodología, el equipo y cómo citar (Ver Figura 2)

Para consultar cada corpus, están disponibles tres tipos de búsqueda: por palabra, también conocida como KWIC (*Key Word In Context*); por metadatos, que en la interfaz general suelen ser dos o tres datos relevantes de acuerdo con el tipo de corpus; y por colocaciones. (Ver Figura 3).

Las consultas aparecerán de acuerdo con la información que se haya ingresado del corpus; por ejemplo, el corpus ASMYCU, hasta el momento, cuenta con los audios sin su transcripción, por lo que se pueden consultar a partir de dos metadatos (temas y lugar de la encuesta). Si a futuro el corpus es transcrito se habilitarán los otros dos tipos de búsqueda, como en el corpus del ALEC.

En cuanto a los resultados, generalmente se muestra la cantidad, el identificador del archivo de la muestra, los resultados y un visor para escuchar y ver la transcripción o el texto. Cuando hay un texto o transcripción, suele ir acompañada en la parte inicial de una ficha de metadatos de la muestra.

Las consultas y resultados que hemos mencionado hasta ahora son las que puede hacer cualquier usuario interesado al ingresar a la plataforma. Sin embargo, en este espacio solo podrá visualizar la información. Para tener otras funcionalidades la persona se puede registrar en el sistema, lo que le permitirá:

- **Hacer consultas especializadas de metadatos:** el usuario registrado suele tener más metadatos disponibles para las consultas. Por ejemplo, en el corpus HCB el usuario registrado, además de los metadatos generales, puede buscar por tema, nivel educativo, o profesión de los informantes.
- **Descargar archivos:** de acuerdo con los permisos del corpus, se pueden descargar muestras o transcripciones en distintos formatos.
- **Visualizar la anotación POS automática de las muestras o transcripciones:** si el corpus tiene transcripción o es de tipo textual, el sistema hace la anotación con los etiquetadores de *Freeling* y *Treetagger*.

Al consultar una muestra el usuario registrado podrá elegir y visualizar la anotación con cualquiera de los dos etiquetadores (Ver Figura 4)

3.5. Seguridad del sistema

Para poder garantizar la seguridad de la información de los corpus, se descartó la opción de crear CLICC en un sistema de gestión de contenidos (CMS). En su lugar, se inició un proceso de desarrollo que siguió todas las fases de la ingeniería de sistemas y se implementó en servidores propios del Instituto para así poder brindar soporte y garantizar la actualización del sistema. Adicionalmente, se implementaron las siguientes medidas:

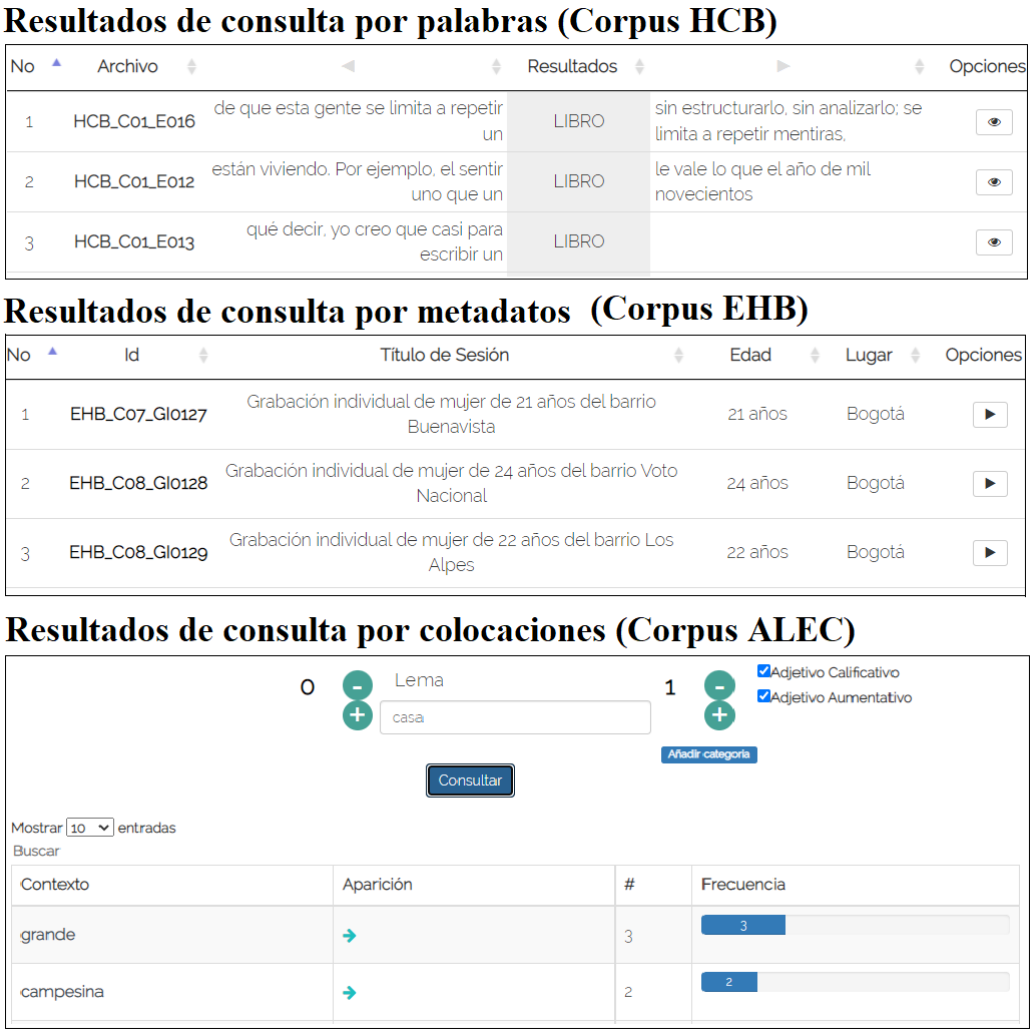

Figura 3: Resultados de consultas en la interfaz de usuario general en CLICC

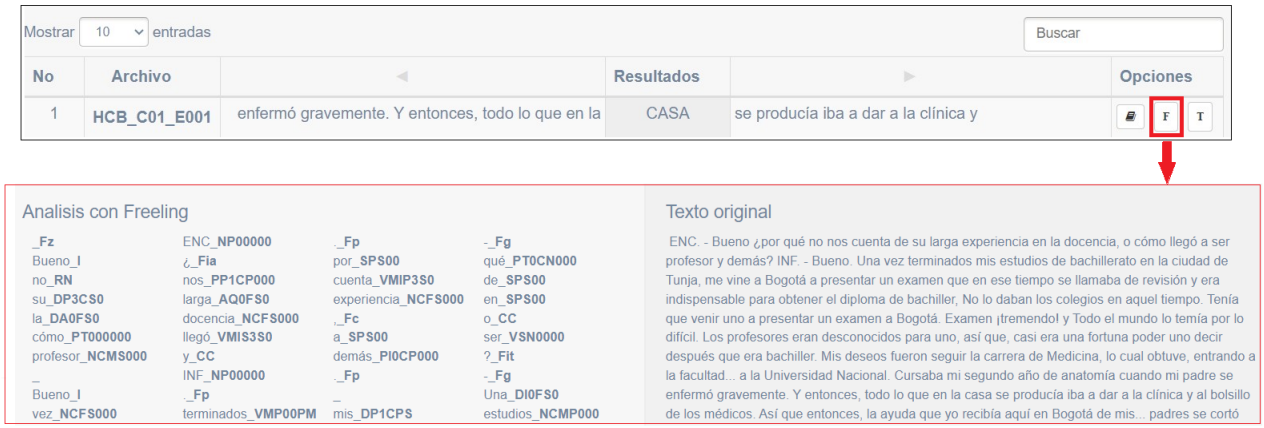

Figura 4: Visualización en IUR de opciones para consulta de etiquetado morfosintáctico y visualización de una muestra con Freeling

- **Contraseñas seguras:** Las contraseñas que se utilizan para acceder a cualquier tipo de usuario de CLICC son creadas por el propio usuario. Para asegurar la protección de la información, se requiere que las contraseñas contengan como mínimo 8 caracteres, combinando mayúsculas y minúsculas.
- **Permisos de usuario:** El sistema se encuentra modularizado, lo que permite la coexistencia de los 4 roles que se pueden asignar a cada usuario y, así, especificar a qué información tiene acceso. Adicional a esto, el sistema permite hacer el seguimiento de las acciones realizadas por cada usuario durante su sesión, ya que cuenta con un registro detallado de las mismas.
- **Verificación de archivos:** El módulo dedicado al cargue de los archivos de los corpus ha sido diseñado para permitir solamente la carga de extensiones de archivo previamente configuradas en el código, de manera que cualquier archivo malicioso será automáticamente rechazado por el servidor y no se permitirá su ingreso al sistema.

3.6. Políticas de seguridad y uso

En CLICC se ha implementado la norma ISO 27001 (ISO, 2022) con el fin de establecer políticas de seguridad robustas y efectivas para proteger los datos almacenados. Entre las políticas implementadas, destacan las siguientes:

- **Identificación y evaluación de riesgos:** Constantemente se monitorea el sistema para identificar cualquier amenaza o vulnerabilidad, lo que permite establecer controles de seguridad adecuados y reducir posibles riesgos.
- **Acceso y control de acceso:** El administrador del sistema es quien tiene la capacidad de permitirlo o rechazarlo. De esta manera, se asegura que solamente el personal autorizado pueda acceder e interactuar con el código, el servidor y la base de datos.
- **Control de acceso físico:** Se aplica una restricción del acceso a los puntos físicos de servidores y centros de datos con el fin de garantizar la protección de los recursos y la información almacenada.
- **Copias de seguridad y recuperación de desastres:** Las copias de seguridad se realizan en el servidor principal ubicado en la sede principal del instituto y, adicionalmente, se crea una copia de respaldo en los servidores alojados en la sede alterna para garantizar la disponibilidad y la recuperación de la información ante posibles contingencias.

Respecto a las políticas de uso, para poder registrarse en CLICC, los usuarios deben aceptar unas condiciones de uso que incluyen compromisos respecto a la confidencialidad del código, el uso de la información con fines no comerciales, entre otros. Para la publicación de los datos, se ha procurado que los corpus cuenten con consentimientos informados de los participantes, y se está avanzando en la definición de mecanismos legales para el cumplimiento de las leyes colombianas de tratamiento de datos y la protección de derechos de autor en la publicación de corpus antiguos y nuevos.

4. Perspectivas futuras

El mejoramiento de la plataforma está orientado actualmente a dos frentes: la adaptación de la plataforma para almacenamiento y consulta de corpus bilingües, y la creación e implementación de un *pipeline* para procesamiento del español. Respecto a los corpus bilingües, se partió de la sistematización de Corpus de Lengua Sáliba, siguiendo los flujos de trabajo con ELAN y FLEx de Gaved & Salffner (2014) y Pennington (2014), y procurando que queden disponibles tanto las transcripciones de audios alineadas en las dos lenguas (sáliba y español), como la información especializada que pueden ofrecer investigadores y sabedores de las lenguas nativas (como la glosa morfológica o la transcripción en diferentes sistemas de escritura). Se están realizando análisis de datos y pruebas con los archivos resultantes de dicha sistematización (txt, eaf y xml) para consolidar un protocolo de ingreso de corpus bilingües y el desarrollo a futuro de una consulta por palabras que ofrezca resultados alineados en las dos lenguas. Por otro lado, se está construyendo un *pipeline* de procesamiento para lematización, etiquetado de partes del discurso y parseado de dependencias, mediante el uso de una arquitectura avanzada de redes neuronales, que sustituirá a *Freeling* en CLICC. Esta tecnología se fundamenta en AnCora (Taulé et al., 2008), como modelo general del idioma español, y en un modelo propio calibrado para el español colombiano, construido a partir de experiencias previas con otros corpus como el Corpus Oral y Sonoro del Español Rural.[5] Estos corpus se están consolidando como referentes para el afinamiento de los modelos de procesamiento del lenguaje natural en español, los cuales suelen ser entrenados con español escrito (Bonilla et al., 2022).

[5] http://www.corpusrural.es

5. Conclusiones

CLICC es una plataforma en línea colaborativa para el almacenamiento, administración y consulta de corpus de las lenguas de Colombia. Cuenta con varios tipos de usuarios que tienen permisos específicos y funcionalidades de acuerdo a sus roles, lo que facilita el trabajo colaborativo y el manejo y privacidad de la información. El sistema está compuesto por una base de datos, una interfaz de usuario general (la web), una interfaz de usuario registrado (para consultas especializadas y descargas), una interfaz de investigador (para el ingreso y administración de corpus) y una interfaz de administración (para la gestión de usuarios y corpus). Hasta el momento se han creado 13 corpus de distintos tipos, de los cuales 6 ya están públicos para consulta en línea y su reutilización en otras investigaciones.

Como perspectivas a futuro, es necesario seguir trabajando en el desarrollo de las funciones para corpus bilingües paralelos y alineados, en la creación e implementación de herramientas especializadas para análisis y procesamiento de muestras del español, y en el mantenimiento y mejoramiento de la plataforma para el tratamiento de diferentes tipos de corpus. Todo con miras a que CLICC se siga consolidando como un espacio para que investigadores, grupos de investigación, comunidades y personas interesadas puedan administrar y publicar sus corpus de las lenguas de Colombia y así contribuir a la documentación, investigación, estudio y promoción de la diversidad lingüística y cultural del país.

Referencias

Anthony, Laurence. 2013. Developing AntConc for a new generation of corpus linguists. En *Corpus Linguistics Conference (Abstracts Book)*, 14–16.

Bonilla, Johnatan, Miriam Bouzouita & Rosa Segundo Díaz. 2022. La construcción del Corpus Oral y Sonoro del Español Rural-Anotado y Parseado (COSER-AP): avances en el etiquetado de partes del discurso. *Revista Internacional de Lingüística Iberoamericana* 20(2). 77–96. 10.31819/rili-2022-204006.

Caminada, Nuno, Violeta Quental & Milena Garrão. 2008. Linguistics Tools: uma plataforma expansível de funções de consulta a corpus. En *WebMedia: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, 364–368. 10.1145/1809980.1810067.

Gaved, Tim & Sophie Salffner. 2014. Working with ELAN and FLEx together: an ELAN-FLEx-ELAN teaching set. https://es.scribd.com/document/357359102/Working-with-ELAN-and-FLEx-together-pdf.

ISO. 2022. ISO/IEC 27001 Information security management systems. Organización Internacional para la Estandarización, https://www.iso.org/standard/27001#lifecycle.

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

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

Pennington, Ryan. 2014. Producing time-aligned interlinear texts: Towards a SayMore–FLEx–ELAN workflow. Unpublised draft: https://www.sil.org/resources/archives/66553.

Rehm, Georg, O. Schonefeld, Andreas Witt, C. Chiarcos & T. Lehmberg. 2008. SPLICR: a sustainability platform for linguistic corpora and resources. En *KONVENS*, 86–96.

Rubio, Ruth & Julio Bernal. 2019. Corpus Oral del Instituto Caro y Cuervo: reestructuración, diseño y construcción. *Lexis* 43(1). 195–219.

Rubio, Ruth, Andrea Llanos, Julio Bernal, Johnatan Bonilla & Daniel Bejarano. 2017. Diseño y elaboración del sistema gestor de contenidos para los corpus lingüísticos del Instituto Caro y Cuervo. En *Estudios Lingüísticos*, Instituto de Literatura y Lingüística. Cuba.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. En *International Conference on New Methods in Language Processing*, 1–9.

Scott, M. 2008. Developing WordSmith. *International Journal of English Studies* 8(1). 95–106.

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

Taulé, Mariona, Antonia Martí & Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. En *Language Resources and Evaluation Conference (LREC)*, s.p.

Torruella, Joan & Joaquim Llisterri. 1999. Diseño de corpus textuales y orales. En *Filología e informática. Nuevas tecnologías en los estudios filológicos*, 45–77. Editorial Milenio.

http://www.linguamatica.com/

linguaMATICA

Artigos de Investigação

Projetos, Apresentam-se