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.

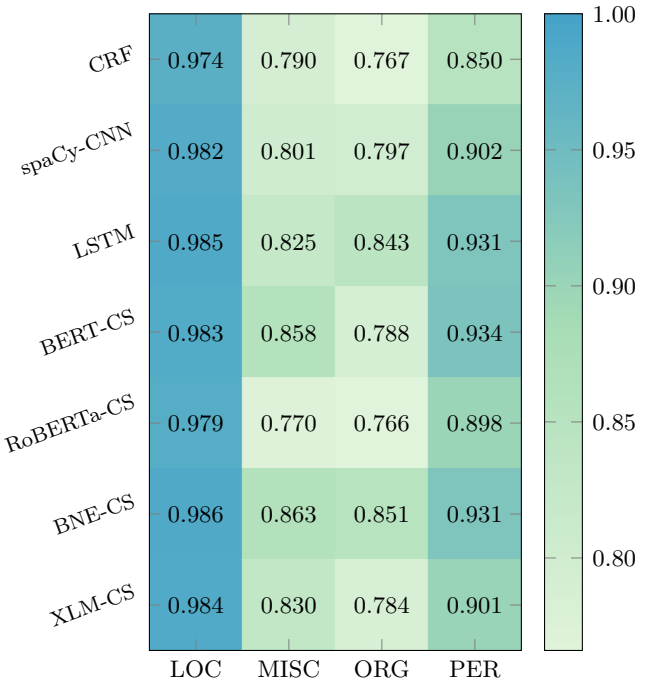

Figura 2: Valores `F1` de todos os experimentos divididos por etiqueta.

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

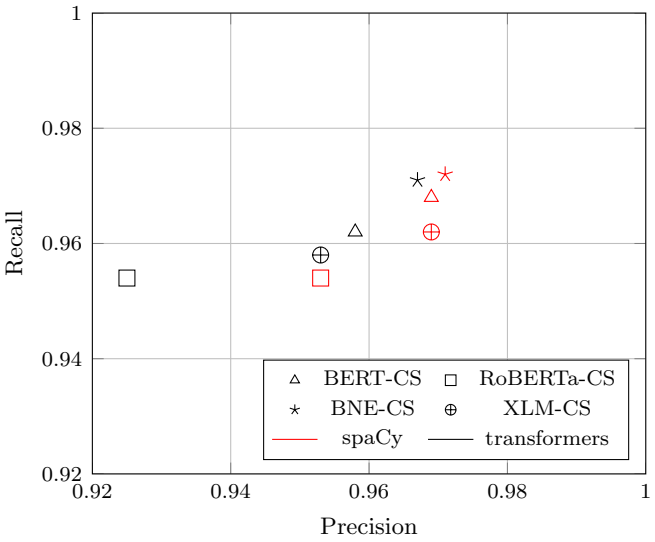

Figura 3: Comparação dos resultados da afinação com as bibliotecas *spaCy* e *transformers*.

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

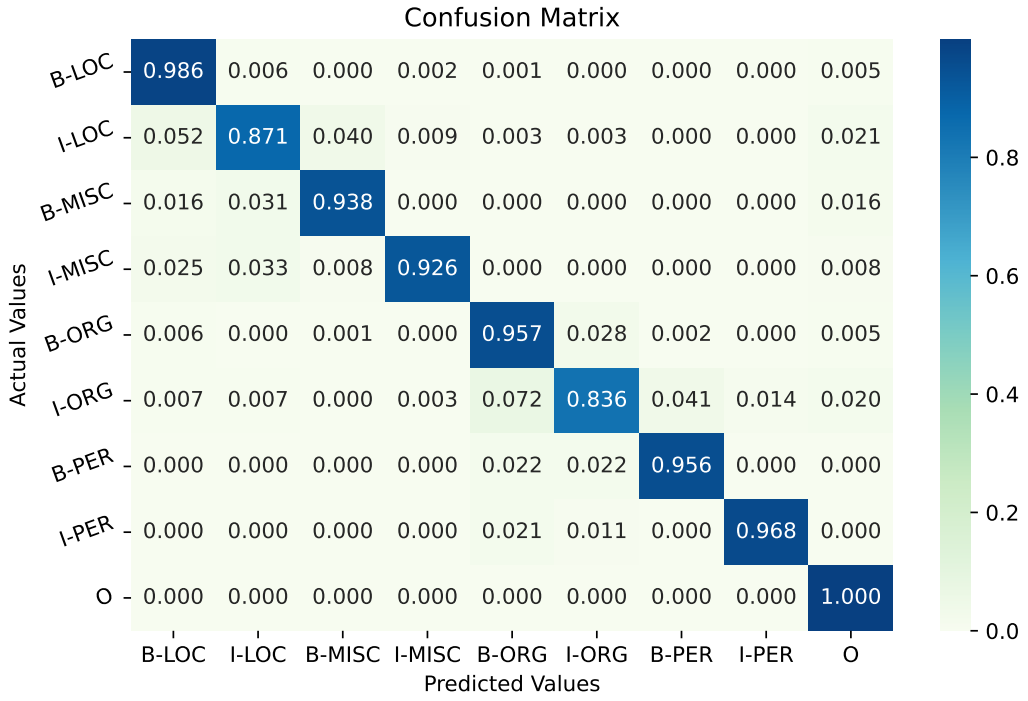

Figura 4: Matriz de confusão do melhor modelo: *BNE-CS* afinado com spaCy.

	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.