Towards the automatic creation of NER systems for new domains

Emanuel Matos and **Mário Rodrigues** and **António Teixeira**
IEETA, DETI, University of Aveiro, Aveiro, Portugal
LASI – Intelligent System Associate Laboratory, Portugal
easm,mjfr,ajst@ua.pt

Abstract

Creation of NER systems for new domains with no annotated data is an unsolved problem. The main objective for this paper is to address some of the limitations of the development of Named Entity Recognition (NER) systems based on Bidirectional Encoder Representations from Transformers (BERT) model using automatically annotated data, making it more suited for new domain scenarios. The proposed extension to the method is based on a state-of-the-art Open Information Extraction (Open IE) system, that combined with a mapper provides automatic annotation to fine-tune Deep Learning (DL) models. A proof-of-concept of the proposal was implemented and assessed with WikiNER dataset. Several factors were studied regarding their influence in the performance: different DL models to serve as basis for the fine-tuning process (BERT, RoBERTa, BART); the number of entities considered; and the training set size. The study confirmed the potential of the approach and demonstrated the capability of models to achieve Precisions above 75% for sets of entities with 20 elements.

1 Introduction

Named Entity Recognition (NER) is an essential natural language processing technique that identifies and categorizes entities in text, such as names of people, organizations, locations, and more. It is relevant for information extraction, entity linking, and various AI applications. NER helps transform unstructured text into structured data, enabling better understanding and utilization of textual information in a wide range of fields and industries.

Recently, there has been a growing interest in enhancing Named Entity Recognition (NER) systems for the Portuguese language. Various techniques have been explored, including Conditional Random Fields (CRF), Long Short-Term Memory networks (LSTMs), and, more notably, Deep Learning approaches since 2020, with the introduction of BERT.

The pioneering utilization of BERT in Portuguese NER, as demonstrated by Souza et al. in their 2020 work (Souza et al., 2020), combined the strengths of BERT with Conditional Random Fields (CRF). This fusion harnessed BERT's transfer learning capabilities while leveraging CRF for accurate entity predictions.

The NER model was trained on the First HAREM dataset and subsequently tested using the MiniHAREM dataset. Remarkably, despite the limited size of the training data, this innovative approach managed to achieve state-of-the-art performance, showcasing the potential of Deep Learning methods even in scenarios with sparse annotated datasets. This success underscores the growing interest in Deep Learning techniques for NER in situations where data resources are constrained.

> "Supervised NER systems, including DL-based NER, require big annotated data in training. However, data annotation remains time consuming and expensive. It is a big challenge for many resource-poor languages and **specific domains** as domain experts are needed to perform annotation tasks." (Li et al., 2022)

When dealing with scenarios such as new domains, where there is access to a small but high-quality annotated dataset, it becomes worthwhile to consider the exploration of bootstrap techniques (Jurafsky and Martin, 2023a). When a small annotated dataset is not even available, alternative solutions become imperative. To tackle this challenge, Matos et al. introduced a solution in their paper Matos et al., 2022a. They proposed the creation of NER systems using BERT, leveraging automatically annotated data. Their approach involved the application of Transfer Learning, fine-tuning pretrained BERT models with a dataset that was automatically annotated and focused on the

Tourism domain, sourced from Wikivoyage texts. The achieved performance was interesting, with the best F1 score reaching 64.9%.

To make this proposal useful in completely new domains several challenges remain:

- Be capable of annotating according to a set of classes that is dependent of the domain.
- Derive that set from existing resources and/or using existing tools.
- Be capable of handling larger sets of classes than the classic ones.

The primary aim of this paper is to introduce an evolution of the approach initially put forward by Matos et al. in their work (Matos et al., 2022a), which was subsequently refined in (Matos et al., 2022c). This enhanced method is tailored to better accommodate new domain scenarios, with a particular focus on addressing the three challenges outlined earlier.

Paper structure – Next section presents relevant related work; section 3 describes the proposed method; Sections 4 and 5 the experimental setup (proof-of-concept) and results obtained.

2 Related Work

"In recent years, DL-based NER models become dominant and achieve state-of-the-art results. Compared to feature-based approaches, deep learning is beneficial in discovering hidden features automatically" (Li et al., 2022).

Language model embeddings pre-trained using Transformer are becoming a new paradigm of NER (Li et al., 2022). These language model embeddings can be further fine-tuned with one additional output layer for NER tasks (Li et al., 2022).

NER tasks are typically structured as sequence labeling problems, where each word in a sequence is assigned a tag. This tagging process is often approached using a multi-layer Perceptron with a Softmax layer as the tag decoder, essentially framing the task as a multi-class classification problem. Each word's tag is predicted independently, solely based on its contextual representations, without considering its neighboring words. Many previously introduced NER models have employed this Multi-Layer Perceptron (MLP) with Softmax as their tag decoder.

Numerous more recent deep learning-based NER models employ a CRF layer as the tag decoder, often in conjunction with bidirectional LSTM or CNN layers (Li et al., 2022).

2.1 Recent evolutions in NER for PT

Table 1 presents recent representative examples of NER for Portuguese, which are briefly described next.

With the aim of recognizing named entities in different textual genres, including genres different from those for which it was trained, Pirovani and collaborators (Pirovani et al., 2019) adopted a hybrid technique combining Conditional Random Fields with a Local Grammar (CRF+LG), which they adapted to various textual genres in Portuguese, according to the task of Recognition of Named Entities in Portugal in IberLEF 2019.

Regarding systems developed for specific contexts, the LeNER-Br system (Luz de Araujo et al., 2018), presented in 2018, was developed for Brazilian legal documents. LSTM-CRF models were trained with Paramopama, obtaining F1 performance of 97.04% and 88.82% for Legislation and judicial entities. According to the authors, the results showed the viability of NER systems for judicial applications.

Lopes et al. (2019) addressed NER for clinical data in Portuguese with BiLSTMs and word embeddings. The performance obtained was an F1 slightly above 80% and equivalent results for Precision and Recall. The dataset was pre-processed by NLPPort (Ferreira et al., 2019) and processed by BiLSTM-CRF and CRF for comparison. BiLSTM was superior in all comparisons for the In-Domain models.

In work published in 2020, NER was applied to the discovery of sensitive data in Portuguese (Dias et al., 2020), being used in the process of protecting sensitive data. A component was developed to extract and classify sensitive data, from unstructured textual information in European Portuguese, combining several techniques (lexical rules, machine learning algorithms and neural networks).

BERT was used for NER for Portuguese in 2020 (Souza et al., 2020). In this work, Portuguese BERT models were trained and a BERT-CRF architecture was used, combining BERT transfer capabilities with structured CRF predictions. BERT pre-training used the brWac corpus, which contains 2.68 billion tokens from 3.53 million documents and is the largest Portuguese open corpus to date. The NER model training was done with First HAREM. Tests on the MiniHAREM dataset out-

Table 1: Recent representative Work in NER for Portuguese.

Ref.	Language	Domain	Technics
(Luz de Araujo et al., 2018)	Brazilian Portuguese	Legal	LSTM-CRF
(Pirovani et al., 2019)	Portuguese	General	CRF+LG
(Lopes et al., 2019)	European Portuguese	Clinical	BiLSTM-CRF
(Dias et al., 2020)	European Portuguese	Sensitive Data	Rule-based, CRF, Random Fields and BiLSTM
(Souza et al., 2020)	Portuguese	HAREM Golden collection	BERT, CRF
(Souza et al., 2023)	Portuguese	HAREM Golden collection	BERT, CRF

performed the previous state of the art (BiLSTM CRF+FlairBBP), despite being trained with much less data.

From the selected representatives of recent NER developments for Portuguese it is clear that: (1) the target domains are quite diverse, being different for all selected references; (2) the set of techniques applied is also diverse, with Machine Learning methods and tools being frequently adopted, including some more recent ones such as LSTM and BERT; (3) NER for Portuguese continues to be a relevant and active area, with developments in line with the evolution of the state of the art; (4) there are signs of expansion of areas/domains of application.

Recent work regarding BERT models for Brazilian Portuguese (Souza et al., 2023), included NER in NLP tasks used for evaluation, with textual sentence similarity, and implication detection. When evaluated in the total scenario (10 entities) with First HAREM and mini HAREM dataset, the BERT model trained for Portuguese NER obtained a maximum Precision and F1 of 78.3% and 75.6%, respectively.

3 Proposed Method

The proposed extension to the method of (Matos et al., 2022b) consists in replacing the NER systems for derivation of automatic annotation by a method more domain agnostic, using Open Information Extraction (OpenIE) methods, and automatically select the set of entities to consider.

The method proposed, represented in Fig. 1, consists of the following main parts:

Domain dataset(s) – that will constitute the input to the process. At this stage of experimental validation of the proposed method, it must include annotations, that are only used for evaluation purposes.

Automatic annotation pipeline – Based in Open Information Extraction (OpenIE), this processing block starts by applying OpenIE to the sentence and, in a second step, mapping to entities the relevant parts of the triples (the ones regarding subject and object). As a final step, before generation of annotated dataset, selection of the top occurring entity classes is made. This top classes will define the entity set to be used in annotation and will be domain dependent.

Fine-tuning of models – Available Deep Learning (DL) models, such as BERT, are fine-tuned to the domain specific entity set, using the train set with automatic annotations resulting from the previous step.

Evaluation in test set – To assess the fine-tuned models, the test set is processed by them and the output compared to the manual annotations. The standard metrics in NER field are produced (Precision, Recall and F1).

4 Proof-of-concept

This section presents how the process outlined in previous section was instantiated to create a first proof-of-concept. Information is given regarding: dataset adopted, automatic annotation process based in OpenIE, entity selection, DL models, and fine-tuning.

4.1 Dataset

For an initial proof-of-concept, as representative of a specific domain, the Portuguese part of the WikiNER dataset was adopted. Created by Nothman (Nothman et al., 2013), the WikiNER dataset contains 7.200 manually labeled Wikipedia articles in nine multilingual languages: English, German, French, Polish, Italian, Spanish, Dutch, Portuguese, and Russian.

The dataset includes manually annotated entities that we only use for evaluation purposes. The set

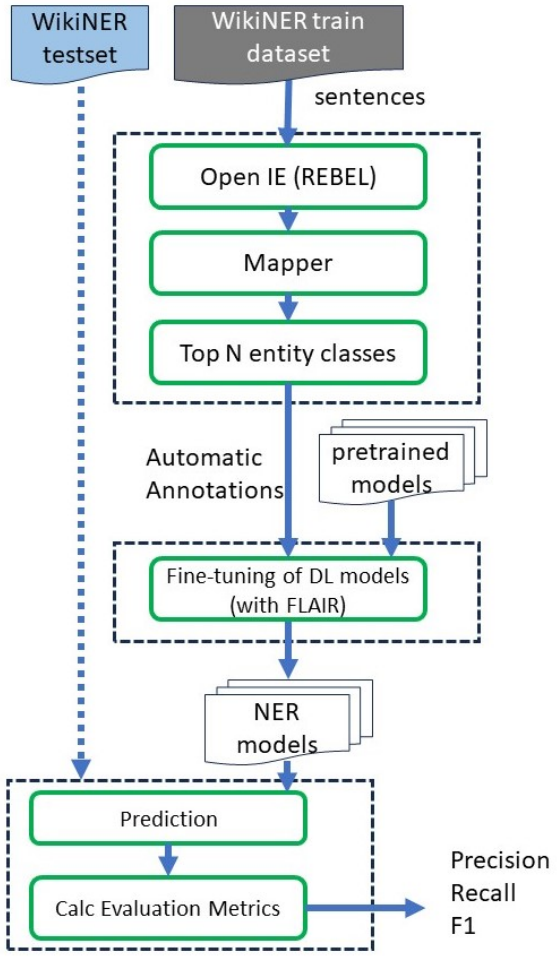

Figure 1: Overall presentation of the method proposed.

of the manually annotated entities is: LOC (e.g., towns); ORG (e.g., musical groups), PER (e.g., living people); MISC (e.g., television series, discographies); NON (e.g., years); DAB (disambiguation).

For the work presented in this paper 142112 sentences from the Portuguese part of the dataset were used (for train, validation and test). Examples of sentences included in the adopted dataset are presented in Table 2.

4.2 OpenIE-based automatic annotation

This block includes OpenIE, mapping to entities and entity selection.

4.2.1 OpenIE

For the proof-of-concept, a representative of state-of-the-art in OpenIE, the REBEL (Relation Extraction By End-to-end Language generation) system (Cabot and Navigli, 2021), was adopted. It is a seq2seq model based on BART that was trained for relation extraction.

O Algarve constitui uma das regiões turísticas mais importantes de Portugal e da Europa.
A Constituição imperial de 1824 tornou o Brasil um país unitário visando facilitar o controle do governo central sobre as províncias e assim impedir um eventual desmembramento territorial.
Aveiro, conhecida como a Veneza portuguesa e durante algum tempo chamada de Nova Bragança, é uma cidade portuguesa, capital do Distrito de Aveiro, na região Centro e pertencente à subregião do Baixo Vouga, com cerca de 55 291 habitantes.

Table 2: Examples of sentences included in the adopted dataset (a subset of WikiNER).

4.2.2 Mapping to entities

To associate an entity tag to a word or sequence of words the following process is applied:

1. The triples extracted by REBEL are processed and a list with only the <obj> or <subj> content is created, keeping information regarding sentence number.

2. Elements in the list obtained in previous step, consisting of a word or sequence of words, are processed by Wikimapper (Klie), one by one, to assign the entity's QID, which is the unique identifier assigned to the entity by Wikidata (Wikimedia Foundation). Each item in Wikidata is assigned a unique identifier called a QID, which is an alphanumeric string. For example, the QID for the Portuguese language in Wikidata is "Q5146", being the information regarding it available at `https://www.wikidata.org/wiki/Q5146`.

   As there is no guarantee that the list element has a corresponding QID, the assignment process makes several tries: first the original words are processed by `wikimapper.title_to_id(word)`. If no QID is returned, the processed is repeated for the singular form of the word(s). If again no QID is returned, translation to English is applied and `title_to_id()` applied to the obtained translation. Examples are presented in Table 3.

3. Get the type for the entity using a query to the Wikidata Query Service (`https://query.wikidata.org/sparql`). The query returns the value of property **P31**, which is used as the type. The property **P31** represents the "instance of" property. It is used to describe

```
<s><triplet> Astrobiologia <subj> advento <obj> studies <subj> sistemas biológicos
    <obj> studies <triplet> advento <subj> Astrobiologia <obj> studied by <triplet>
    sistemas biológicos <subj> Astrobiologia <obj> studied by</s>
<s><triplet> Pólo Sul <subj> nível do mar <obj> tributary <triplet> nível do mar <
    subj> Pólo Sul <obj> mouth of the watercourse</s>
<s><triplet> América do Sul <subj> Atlântico <obj> located in the administrative
    territorial entity <triplet> Atlântico <subj> América do Sul <obj> contains
    administrative territorial entity</s>
<s><triplet> América do Sul <subj> áreas litorâneas <obj> instance of</s>
<s><triplet> povoar <subj> colonizar <obj> has part <triplet> colonizar <subj>
    povoar <obj> part of</s>
<s><triplet> México <subj> continente americano <obj> continent</s>
<s><triplet> civilização Inca <subj> América do Sul <obj> located in the
    administrative territorial entity</s>
```

Figure 2: Example of output from REBEL processing, showing the tags added to mark the triplets and their parts.

Table 3: Examples - Wikimapper Output

Line	Word	ID	Lang	Mapped object
1	Bíblia	Q1845	pt	Bíblia
1	Cosmologia_Bíblica	Q2566489	pt	Cosmologia_Bíblica
2	1048	Q19359	pt	1048
2	Omar_Khayyam	Q35900	pt	Omar_Khayyam
4	Espectro	Q16608018	pt	Espectro
5	Carregadas	Q413088	en	Loaded
9	Estados_Unidos	Q30	pt	Estados_Unidos

the type or class that an item belongs to and is one of the most fundamental properties in Wikidata and is used to categorize items by specifying what kind of thing they are.

4. Use the type obtained from wikidata as the tag for the word (or sequence of words;

4.2.3 Entity selection

Based on occurrence statistics associated to each entity type (tag), 4 different datasets were created keeping only the N tags with higher occurrences, with $N = 5, 10, 15$ and 20. The output was saved in BIO format. The lists of automatically derived sets are presented in Table 4.

4.3 Fine-tuning of DL Models

We selected as tool for our experiments the state-of-the-art deep learning framework FLAIR (Akbik et al., 2019), designed keeping in mind the ease of parameter tuning and implementation while training using any embedding model on the dataset, characteristics essential for our objectives. This framework is commonly used in information extraction tasks, such as NER and Relation extraction. It provides a unified interface for word embeddings and flexibility in combining multiple embeddings, known as stacked embeddings.

For this work, the fine-tuned models for NER applied to our domain/task were the following:

ner-bert (BERT): `bert-base-pt-cased`[1] model (Abdaoui et al., 2020), a smaller version of `bert-base-multilingual-cased` for Portuguese. It was obtained by breaking the multilingual transformers into smaller models according to the targeted languages. Was selected due to its smaller size.

ner-roberta (ROBERTA): `xlm-roberta-base-trimmed-pt-60000`[2] model based in `xlm-roberta-base`, pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. It was introduced in the paper Unsupervised Cross-lingual Representation Learning at Scale (Conneau et al., 2020).

ner-bart (BART): `bart-large-mnli`[3] a checkpoint for `bart-large` after being trained on the MultiNLI (MNLI) dataset. BART is a transformer encoder-decoder (seq2seq) model

[1] https://huggingface.co/Geotrend/bert-base-pt-cased

[2] https://huggingface.co/vocabtrimmer/xlm-roberta-base-trimmed-pt-60000

[3] https://huggingface.co/facebook/bart-large-mnli

Table 4: Information regarding the sets of entities automatically derived. N represents the number of top occurring entities selected.

N	Entities
5	['país', 'ser_humano', 'cidade', 'ano', 'município_do_Brasil']
10	+ ['capital', 'unidade_federativa_do_Brasil', 'estado_dos_Estados_Unidos', 'Estado_soberano', 'município_de_Portugal']
15	+ ['profissão', 'continente', 'freguesia_de_Portugal', 'táxon', 'especialidade']
20	+ ['designação_para_uma_entidade_territorial_administrativa_de_um_país_específico', 'género_musical', 'banda_musical', 'ilha', 'banda_de_rock']

with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is particularly effective when fine-tuned for text generation (e.g. summarization, translation) but also works well for other tasks (e.g. question answering) (Lewis et al., 2019).

As at least part of the entities in Portuguese use capital words cased models were adopted. Also, to take into account the context, CRF was adopted for the output layer.

4.3.1 Fine-tuning

Training (and evaluation) of the models was performed in 2 computers with GPUs. Details of the configurations are presented in table 5.

Before starting tests with the algorithms and our hypotheses, we trained the 3 different models (`bert-base-pt-cased`, `xlm-roberta-base-trimmed-pt-60000` and `bart-large-mnli`) using 50 epochs. The variation of F1 and loss, in Fig. 3, showed stabilization or inversion of the descent (for loss) around the 10th epoch. Therefore, 10 epochs was adopted as the training stop criteria for the all the experiments.

5 Results

Examples of annotations obtained with the trained models are presented in Table 6. Next subsections present the commonly used metrics (Precision, Recall and F1) and how they are affected by relevant factors: DL model, training set size, number and type of entities.

5.1 Effect of the DL model

The results as function of model considered are presented in Fig. 4. Similar information adding number of entities as a factor is presented in Fig. 5.

Figures 4 and 5 show values around 70, 50 and 60 for Precision, Recall and F1, respectively. The results don't differ much across models, but results are worst for BART. This is more noticeable in

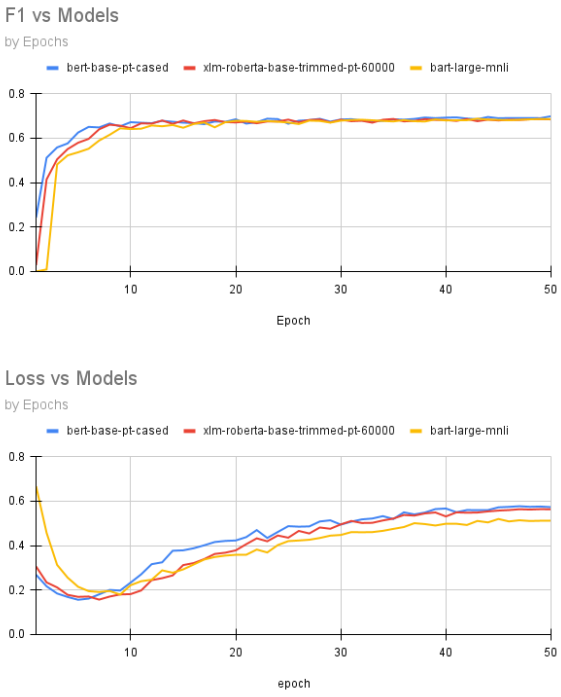

Figure 3: F1 and Loss vs Models by Epochs.

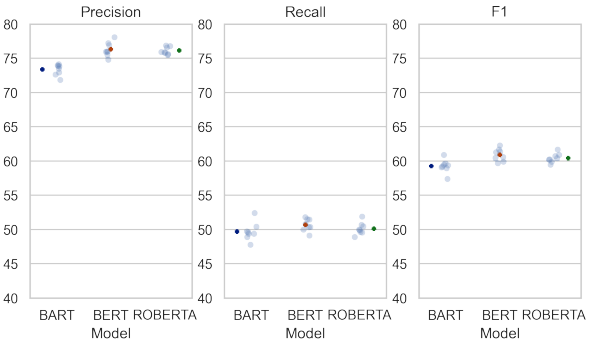

Figure 4: Results as function of Model, considering all entities. The crosses represent the mean value.

Precision. Also, the number of entities considered does not seem to affect much the results.

5.2 Effect of train set size

To investigate possible effect of train set size, the results for each of the 2 train sizes used are presented, separately, in Fig. 6.

Table 5: Details Notebooks

Notebook	GPU RAM	CPU RAM	OS	Chipset
Apple	24	32	Sonoma 14.0	Apple M1 Max
Asus	8	16	Windows 11	Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz

Table 6: Examples of annotations obtained with the systems developed.

```
O oeste e sul da Áustria/B_PAÍS estão situados nos Alpes/B_OTHER, o que faz do ↩
país um destino bem conhecido de desportos de inverno.

Em 1874, mil anos após o estabelecimento da colónia de Ingólfur/B_OTHER ↩
Arnarson, a Dinamarca/B_PAÍS concedeu à Islândia/B_PAÍS autoridade interna, que ↩
foi renovada em 1904/B_OTHER.
```

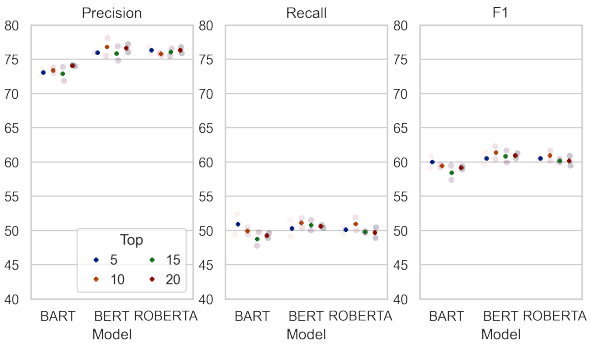

Figure 5: Results considering all tags as function of Model and number of entities (Top). The crosses represent the mean value.

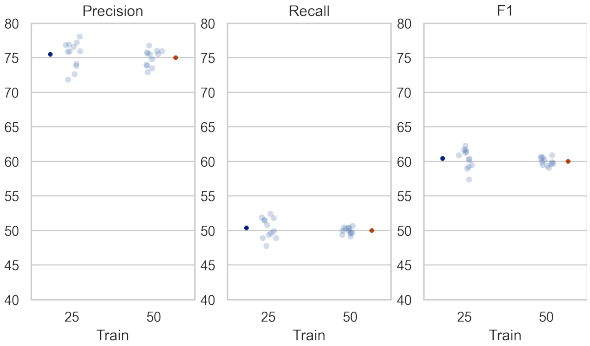

Figure 6: Results considering all tags as function of Train size). The crosses represent the mean value.

The plots show very similar results for the two train sizes used, for all 3 metrics. No advantage of a larger train set was found in the results.

5.3 Effect of number and type of entities

As it is very relevant to assess if the models are capable of handling different sizes of entities' sets, the results as function of number of top occurring entities considered are presented in Fig. 7.

The results obtained, with average values for

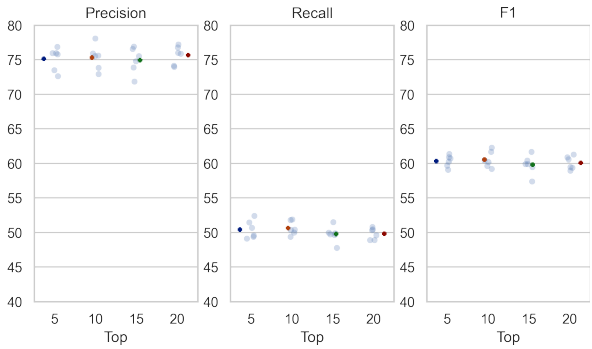

Figure 7: Results as function of number of entities, considering all tags. The crosses represent the mean.

Precision, Recall and F1 very similar, indicate that models are capable of maintaining the performance for all the sets considered, including the larger one, with 20 entities.

Complementing the information in Fig. 7, the precision for each of the entity types is presented using stripplots in Fig. 8. Results are presented separately for each size of the entities set considered in the experiments.

The plots show that there are several entities with high precision, close to 100 %; the entity "OTHER" despite its different nature attains precision around 80%; there are types, such as "ANO" (year) that the system is not good at; with the increase in number of entity types (and reduction of number of examples in train set) more entities with low precision appear, as, for example, "ESPECIALIDADE"(specialty).

5.4 Generalization capability of the models

To conclude the analyses, a very preliminary analysis of the "learning" capabilities of the models was performed. For this, the words of the test set

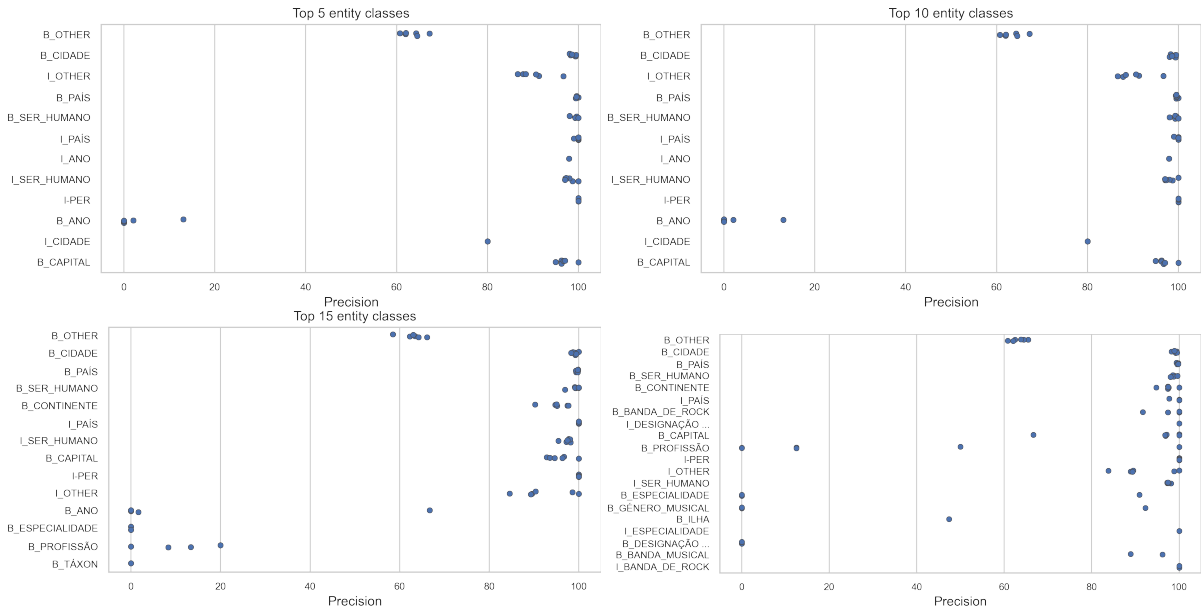

Figure 8: Precision obtained for the several entity classes as function of the number of top occuring entities retained (5,10, 15 and 20).

tagged as entities and not present in the train set were obtained. The number of novel words in test set, for a train set of 25 %, are presented in Table 7. Due to space limitations are presented only results for the smaller and larger set of entities considered in this study.

Table 7: Statistics regarding the number of words not in the training set annotated by the systems developed. Values are presented for the 3 models and two sizes of the set of entities (5 and 20).

Model	Top	Novel words	Annotated	% An.
BART	5	1083	2090	51.8
BERT	5	996	1944	51.2
ROBERTA	5	934	1855	50.4
BART	20	940	1881	50.0
BERT	20	949	1889	50.2
ROBERTA	20	942	1887	49.9

The novel annotated words represent approximately 50% of the annotated words, being the highest number of novel words 1083 (51.8%), obtained when using BART. A fragment of the word list obtained for this case is presented in Table 8. Most of them make sense as entities.

Table 8: Fragment of the 1083 words annotated by the BART model as entity and not present in the train set.

Hatshepsut, monazita, Mônica, druida, etnia, Leeds, Estandarte, Ismênia, sátiros, Honolulu, Etti, Nasceu, arcades, agrotóxicos, Mario, Guam, Portas, Barbosa, Amazon, Memórias, proletariado, magiares, 2002, o, McLaren, ...

6 Conclusion

Addresssing the challenge of creation of NER systems for new domains with no annotated data, this paper proposed the use of an OpenIE-based NER to provide automatic annotation to support fine-tuning of state-of-the-art DL models for NER in Portuguese. Experiments were performed with 3 DL base models, different numbers of entities, and different train sizes. The values obtained for Precision, around 75%, even for a set of 20 entities, not far from the 78.3% of (Souza et al., 2023). The metrics obtained can be considered a lower bound as many of the annotations considered False Positives are due to not being manually annotated despite being good candidates for consideration as entities. Interesting results were obtained regarding the capability of the trained models to annotate words not present in the training set, pointing to good generalization capacity.

6.1 Future work

The results point to the potential of the approach but many challenges an limitations remain. Future work should include: (1) improvements to the automatic annotation pipeline, starting by adaptation of REBEL to Portuguese, but also contemplating improvement in entity assignment (e. g., adding additional step to obtain entity type when wikidata queries fail); (2) experimentation with several stages of fine tuning. The initial train with part of the train set could be continued using other parts of

the dataset; (3) integration of the best performing models into an ensemble of NER systems such the one created by (Matos et al., 2021); (4) exploration of span-based approaches to NER (Jurafsky and Martin, 2023b) (5) exploration of the potential of bootstrapping methods; (6) adoption of methods to improve balance of the dataset regarding examples in train (and test) set for each type of entity; (7) exploration of recent DL models such as GPT-3 or FLAN (Wei et al., 2021; Brown et al., 2020).

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutlilingual bert. In *SustaiNLP / EMNLP*.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Mariana Dias, João Boné, João C Ferreira, Ricardo Ribeiro, and Rui Maia. 2020. Named entity recognition for sensitive data discovery in portuguese. *Applied Sciences*, 10(7):2303.

João Ferreira, Hugo Gonçalo Oliveira, and Ricardo Rodrigues. 2019. Improving NLTK for processing portuguese. In *8th Symposium on Languages, Applications and Technologies (SLATE)*.

Daniel Jurafsky and James H. Martin. 2023a. *Speech and Language Processing*, chapter 21 - Relation and Event Extraction. Draft of January 7.

Daniel Jurafsky and James H. Martin. 2023b. *Speech and Language Processing*, chapter 11 - Fine-tuning and Masked Language Models. Draft of January 7.

Jan-Christoph Klie. wikimapper. https://github.com/jcklie/wikimapper, Accessed 5 nov 2023.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. 2019. Contributions to clinical named entity recognition in portuguese. In *Proc. 18th BioNLP Workshop and Shared Task*.

Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *PROPOR*, LNCS. Springer.

Emanuel Matos, Mário Rodrigues, Pedro Miguel, and António Teixeira. 2021. Towards Automatic Creation of Annotations to Foster Development of Named Entity Recognizers. In *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*, volume 94 of *OASIcs*, pages 11:1–11:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Emanuel Matos, Mário Rodrigues, Pedro Miguel, and António Teixeira. 2022a. Named Entity Extractors for New Domains by Transfer Learning with Automatically Annotated Data. In *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 288–298. Springer. https://link.springer.com/chapter/10.1007/978-3-030-98305-5_27.

Emanuel Matos, Mário Rodrigues, and António Teixeira. 2022b. Named entity extractors for new domains by transfer learning with automatically annotated data. In *International Conference on Computational Processing of the Portuguese Language*, pages 288–298. Springer.

Emanuel Matos, Mário Rodrigues, and António Teixeira. 2022c. Assessing Transfer Learning and automatically annotated data in the development of Named Entity Recognizers for new domains . In *Proc. IberSPEECH 2022*, pages 191–195.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Juliana PC Pirovani, James Alves, Marcos Spalenza, Wesley Silva, Cristiano da Silveira Colombo, and Elias Oliveira. 2019. Adapting NER (CRF+ LG) for many textual genres. In *IberLEF@ SEPLN*, pages 421–433.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Portuguese Named Entity Recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2023. BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing*, page 110901.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wikimedia Foundation. Wikidata. https://www.wikidata.org, Accessed 5 nov 2023.