Language Resources and Evaluation (2025) 59:665–684
https://doi.org/10.1007/s10579-024-09738-8

ORIGINAL PAPER

Data augmentation and transfer learning for cross-lingual Named Entity Recognition in the biomedical domain

Brayan Stiven Lancheros[1] · **Gloria Corpas Pastor**[2] · **Ruslan Mitkov**[3]

Accepted: 5 April 2024 / Published online: 10 May 2024

Abstract

Given the increase in production of data for the biomedical field and the unstoppable growth of the internet, the need for Information Extraction (IE) techniques has skyrocketed. Named Entity Recognition (NER) is one of such IE tasks useful for professionals in different areas. There are several settings where biomedical NER is needed, for instance, extraction and analysis of biomedical literature, relation extraction, organisation of biomedical documents, and knowledge-base completion. However, the computational treatment of entities in the biomedical domain has faced a number of challenges including its high cost of annotation, ambiguity, and lack of biomedical NER datasets in languages other than English. These difficulties have hampered data development, affecting both the domain itself and its multilingual coverage. The purpose of this study is to overcome the scarcity of biomedical data for NER in Spanish, for which only two datasets exist, by developing a robust bilingual NER model. Inspired by back-translation, this paper leverages the progress in Neural Machine Translation (NMT) to create a synthetic version of the Colorado Richly Annotated Full-Text (CRAFT) dataset in Spanish. Additionally, a new CRAFT dataset is constructed by replacing 20% of the entities in the original dataset generating a new augmented dataset. We evaluate two training methods: concatenation of datasets and continuous training to assess the transfer learning capabilities of transformers using the newly obtained datasets. The best performing NER system in the development set achieved an F-1 score of 86.39%. The novel methodology proposed in this paper presents the first bilingual NER system and it has the potential to improve applications across under-resourced languages.

Keywords Biomedical NER · Named entity recognition · Spanish · Data augmentation

Extended author information available on the last page of the article

1 Introduction

In the field of Natural Language Processing, Named Entity Recognition (NER) is not a new notion. Since its introduction by the MUC-6 in 1995, it has been a subtask of Information Extraction with substantial research reported. Named Entities (NEs) are textual items of interest with people, organisations, locations, and numbers being common examples. The aim of NER is to recognise and categorise different types of named entities in a structured or unstructured text. This domain has attracted a lot of interest. Starting with rule-based systems (Appelt et al., 1995; Weischedel, 1995) which needed a group of experts creating hand-written rules such as orthographic patterns, and followed by machine learning models (Bam & Shahi, 2014; Borthwick, 1999), and deep learning models (Lin et al. 2017; Chiu & Nichols, 2016; Devlin et al. 2019), the development of systems capable of recognising these NEs has been considerable.

In the general domain, state-of-the art models can produce excellent results. However, research in specialised domains involving different NE classes has not grown at the same rate. The biomedical domain, for example, is expanding as medical records become more computerised and online biomedical research becomes more accessible. According to Microsoft (n.d.), PubMed adds two biomedical papers every minute, thousands every day, and over a million every year. NER is critical for Natural Language Processing (NLP) since NEs serve as both a referential base for finding information in texts and as important pieces of information. A news article could be summarised in the general domain by extracting the five Ws (who, what, when, where, and why). Each W often corresponds to a NE (Zhang et al., 2004, p. 1). Similarly, the biomedical domain makes use of NER since denominations for genes, proteins, and diseases, among other things, are crucial bits of information for researchers and biomedical experts in situations like literature-based discovery and relation extraction. Because biomedical information is primarily published in English, data in other languages, especially NER datasets, are sparse. To fill in this gap, this research will try to overcome the lack of data in other languages by investigating and analysing the best opportunities for developing a bilingual model that can be used in both Spanish and English.

The available datasets for training NER and other NLP systems appear to be insufficient to handle all new information as the diversity and size of online data grows. The cost of manually annotating data for the biomedical domain is high, due to its level of speciality. Moreover, biomedical NER faces challenges such as variations in spelling, synonyms, and unknown vocabulary, which slows the development of new systems. In this paper we explore two data augmentation techniques to increase the size of the dataset: (a) translation of the dataset using a commercial machine translation system to create a dataset in another language; and (b) entity replacement in which a new dataset is constructed by replacing part of the entities in the original dataset (Liu et al., 2020).

Additionally, in order to create a cross-lingual NER model, we propose the use of Transfer Learning, which is the process of training a model using previously learnt parameters from a pre-trained model (Hira et al., 2019), i.e., use the

parameters of a pre-trained model in the *X* domain to initialise a model in the *Y* domain. Continuous training is defined as the sequential training of a system using previously obtained knowledge from one or more data sources. Transfer learning is commonly used to fine-tune general models on new domains or languages with promising results using transformers. Saunders et al. (2019) employed this strategy to test in three domains: biological and health, as well as a general biomedical corpus. They were able to demonstrate an increase in the BLEU score in the combination of biological + biomed for English of + 7 points by first training a base model and then finetuning to the other two domains.

The original methodology that we put forward in this paper will make it possible for additional languages to benefit from biomedical datasets for NER. Our novel approach is portable to other languages and to the best of our knowledge, has not been proposed in the biomedical domain.

The following are this work's main contributions:

- The creation of a synthetic version of the CRAFT corpus in Spanish for the biomedical domain using a cheap translation approach based on back-translation. Using entity replacement, a separate version of the original CRAFT in English was also produced.
- The first bilingual NER system (ES-EN) for the biomedical area, which achieved the second highest F1 score compared to the literature's reports for systems trained in the monolingual CRAFT dataset.

The rest of the paper is structured as follows: Section 2 surveys related work in the field and Sect. 3 presents our methodology, providing details on the experiments conducted. Section 4 discusses the evaluation results and finally Sect. 5 summarises the conclusions of this research.

2 Related work

Although biomedical NER is not a new concept, there is not a global agreement on the entity classes or the annotation criteria, which has led to a handful of datasets with different entity classes and different annotation guidelines, creating inconsistency in the task. For instance, chemical entities are represented in the CHEMDER dataset, which includes the tags: Abbreviation, Family, Formula, Identifier, Multiple, No Class, Systematic, and Trivial (Krallinger et al., 2015). Another chemical dataset is the BC5CDR which also includes diseases (Li et al., 2016). NCBI is a full disease dataset created from 793 PubMed abstracts (Doğan et al., 2014). For proteins/genes identification, the GENIA dataset has a total of 23,996 tagged genes/proteins (Tanabe et al., 2005). The CRAFT dataset also contains genes and proteins but is tagged using instructions from ontologies which might differ from the annotation of other datasets. By combining the datasets provided by the MEDIQA challenge with a subset of the MedQuAD dataset, Lamurias & Couto (2019) presented research on data augmentation for datasets for the Question Answering task, reporting an increase of

0.015 in accuracy in the test set and a decrease of 0.02 in accuracy in the development set.

The Spanish language has not seen the same progress in annotation. The first biomedical NER task in Spanish was called PharmaCoNER, using a chemical dataset containing four tags: No normalizables, Normalizables, Proteinas and Unclear (Gonzalez-Agirre et al., 2019). Focusing on cancer vocabulary, Miranda-Escalada et al. (2020) created the CANTEMIST task for named entity recognition. This corpus contains the following entity types: Disease, Drugs, Unit of Measurement, Excipient, Chemical Composition, Pharmaceutical Form, Medicament, Food, Route, and Therapeutic Action, with a total of 2241 entities.

The lack of standard global guidelines has impeded a unified effort such as that seen in the general domain. As shown above, the entity types differ greatly in all datasets in Spanish or English. Since the objective of this study is to build a single efficient bilingual NER system, the use of a dataset of each language for a single model is not viable, as might be the case for the general domain using the CONLL task's dataset for Spanish and English. Rather, this would bring inconsistency and poor performance to the NER system.

Early Named Entity Recognition systems utilised rule-based methods in which techniques such as transducers (Appelt et al., 1995), pattern matching on dictionaries (Grishman, 1995), and lexical pattern matching (Weischedel, 1995) were used obtaining as high as 94 F-1 scores. Although these systems performed well, considerable human effort was required for writing the rules and maintaining them. These systems were followed by machine learning-based systems that used supervised methods. For instance, an HMM model worked for the entity classes of proteins, RNA, DNA, and cells (Ponomareva et al., 2007), while a semi-Markov model demonstrated the capacity of such models to integrate information across all tokens in a segment (Leaman & Lu, 2016). Unsupervised methods required less manually annotated data. Representative research include "KNOWITALL" by Etzioni et al. (2005) using bootstrapping which leverages little annotated data to learn patterns, a model based on gazetteer creation (Nadeau et al., 2006), and Habib and van Keulen (2012) who presented a method that uses a lookup strategy against knowledge bases like Yago and DBPedia, as well as an unsupervised method for disambiguation, showing the usefulness of online resources for obtaining and extracting information.

Although these early systems achieved competitive performance in the general domain, they were unable to overcome the dynamics of language and frequently performed poorly on unseen data. To tackle this, deep learning methods based on neural networks (NNs) were employed. Such NNs could learn character and word-level representations, as well as sentence-level relationships, in their learning pipeline. Neural Network models tested Convolutional Neural Networks (CNNs) plus character embeddings and obtained an F1 score of 76.39, while a bi-directional long short-term memory (BiLSTM) character embedding approach achieved a 76.94 F1 score for a diseases dataset (Sahu & Anand, 2016). As with the general domain, BiLSTM-CRF (Conditional Random Field) models are popular in the biomedical domain (Cho & Lee, 2019; Li et al., 2019; Wang et al., 2019) with different combinations of character and word embeddings and features such as POS tags.

Contextualised word representation based on transformers is the state of the art for many NLP tasks, and biomedical NER is no exception. Lee et al. (2020) introduced BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) based on BERT. This transformer was trained on PubMed abstracts and PMC full-text articles. Beltagy et al. (2019) re-trained BERT using biomedical data from Semantic Scholar to create SciBERT. The training corpus included 1.14 million full papers. Lastly, Carrino et al. (2021) trained a transformer-based system for Spanish using data from Scielo, Wikipedia, patents, EMEA, and PubMed, among others. This model is based on a RoBERTa-based transformer. Jofche et al. (2022) present a platform for recognising pharmaceutical documents and performing NER and coreference resolution. This platform uses transformer-based architecture to identify NEs in the BC5CDR and BioNLP15CG datasets in English.

Biomedical NER faces a host of issues: the style and speciality of biomedical data, ambiguity, as some names can refer to one entity class or the other, and the constant discovery and coining of new terms, which creates the issue of unknown words, among other challenges. For Zhao et al., (2021, p. 5) the difficulties faced when dealing with biomedical NER are that the biomedical terms have many variations: (1) small variations such as typos, hyphens, or capitalisation, e.g., 'FOXP2' and 'FOX-P2', (2) synonyms and abbreviations, and (3) unseen entities. Additionally, for Cho and Lee (2019, p. 5), the difficulties are the entity boundaries, compound noun phrases, bracket-enclosed entities, nested entities, and the corpus annotation inconsistency. These variations of the NEs complicate the recognition and normalisation of such entities. Another challenge in biomedical NER is the so-called "long-tail" Nes, i.e. "named entities that are rare, often relevant only in specific knowledge domains, yet important for retrieval and exploration purposes." (Liu et al. 2020b, p. 79).

On the basis of the above discussion, one might conclude that the biomedical NER is far from solved, even with state-of-the-art NLP systems. Efficient NLP techniques are needed for automating the extraction and analysis of biomedical data and would help synchronise the efforts across languages.

Previous studies have tested the ability of transformer systems to generalise enough and provide competitive results when trained in another language. Sun and Yang (2019) tested the usability of mBERT (without biomedical data training) and BioBERT (without training in data in Spanish) for the PharmaCoNER dataset. Their results reported competitive F1 scores in both systems (89.24 and 89.02 respectively), as well as the success of current systems in zero-shot transfer. Hakala and Pyysalo (2019) presented an approach of using mBERT for Spanish biomedical named entity recognition without further training, achieving an F1 score of 87 in the test set on the PharmaCoNER dataset.

Mueller et al. (2020) used diverse data in multiple languages to examine the polyglot capabilities of transformers. They show how multilingual models share a large number of parameters, how language-specific training uses those common parameters, and how these models preserve the top 5% of weights for each language. This work reveals that transfer learning is available to create cross-lingual systems in NLP. Saunders et al. (2019) also tested transfer learning for cross-lingual Machine

Translation (MT) models, improving the BLEU score by 7 points, starting with a general domain MT system and training it into a biomedical MT system.

Mayhew et al. (2017) used dictionaries to translate an English dataset into several languages and language resources such as PANLEX and created a phrase translation table in which the labels are copied from source to target sentence. They also tested translating a dataset with Google translate but reported difficulties with the alignment and projection of entities into the new language, which according to the authors resulted in a noisy dataset with incorrect entity tags. Finally, Li et al. (2020) created a model that labels a bilingual corpus and processes NER, then uses GIZA++ for the alignment and extracts NE translation pairs, which are then ranked by calculating the mutual information (MI) value.

3 Methodology and experiments

The selection of the dataset(s) to be used is one of the most important factors to consider when training a system for NER (or any other task). A dataset with a wider coverage of NEs would be more beneficial to Spanish users, especially if the NEs it contains have never been explored before. In the general domain there are datasets for English and Spanish that have the same number of entity tags, the same annotation guideline, and the same class names, such as the CONLL NER task. This is not the case in the biomedical domain, where most datasets only contain a small number of NE classes based on the task at hand. As previously stated, the use of a dataset from the same family with golden annotation is ruled out.

This study will use the CRAFT[1] corpus to train the NER systems as it is one of the datasets with a larger number of entity tags. The CRAFT corpus is a collection of 97 full-text articles from the biomedical domain, manually annotated (version 1 includes 67 articles), and around 100,000 annotations of nine ontologies: Cell type ontology (CL), Chemical entities of biological interest (CHEBI) ontology, NBCI taxonomy (Taxon), Protein Ontology (PR), Sequence Ontology (SO), entries of the Entrez gene database, and three subontologies from the Gene Ontology (GO) (Bada et al., 2012). Our goal is to develop a system able to transfer the knowledge from one of the biggest and most diverse datasets for biomedical NLP from English to Spanish.

The entities contained in the CRAFT corpus are described by Bada et al. (ibid. pp. 11–17) as:

- Chemical: Chemical Entities of Biological Interest refer to atoms, biomedical roles and applications, subatomic particles, molecules, and polyatomic entities.
- Cell: All cell mentions except the types of cell line cells.
- Gene: Biological processes, including at the level of molecules, and subcellular structures. Also: cellular components representing subcellular structures, both intracellular and extracellular; and macromolecular complexes.

[1] http://bionlp-corpora.sourceforge.net/CRAFT/

– Taxon: Biological taxonomy and their corresponding organisms.
– Protein: Based on the protein oncology without regard to sequence type.
– Sequence: Biomacromolecular sequences, attributes, and processes.

We have normalised the annotation in the dataset, meaning that instead of three-letter codes for entities we used the full name of the entity: Protein, Cell, Taxon, Sequence, Chemical, and Gene. This dataset uses the IOB annotation scheme: B-beginning, I- inside and O for tokens not corresponding to any tag, e.g., B-Protein, and I-Taxon. The dataset is divided into three parts: a training set (10,875 elements), a test set (7425 elements), and a validation set (3730 elements).

One important point to note is that the CRAFT dataset does not have a Spanish equivalent. Given the scarcity of datasets in Spanish for the biomedical domain, we would like to offer new ways to improve biomedical systems for Spanish users. As a result, to boost the dataset's size, we used data augmentation techniques.

3.1 Data augmentation

Data augmentation is a strategy to increase a system's performance by generating more training data. Two strategies for the construction of the utilised datasets are presented next: Cheap translation and Entity replacement.

3.1.1 Cheap translation

This initial data augmentation strategy used to train a bilingual NER system was inspired by the success of back-translation in NMT, which is the production of a synthetic dataset via machine translation. According to Edunov et al., (2018, p. 1) "The result is a parallel corpus where the source side is synthetic machine translation output while the target is genuine text written by humans." This method has been shown to improve MT system performance and introduce previously unseen words and entities. Authors also report "that synthetic data can achieve up to 83% of the performance attainable with real bitext" especially if the data matches the domain of the model (Edunov et al., ibid, p. 497).

Google Translate is one of the most widely used machine translation technologies in the world. It supports over 100 languages and its mobile app has been downloaded over 1 billion times (Pitman, 2021). Caswell and Liang (2020) reported a change to the transformer architecture on Google's MT system, which resulted in a +5 BLEU score increase in high-resource languages and a +7 BLEU score increase in low-resource languages on average. This system was chosen to translate the CRAFT dataset into Spanish because of its constant development, widespread use, and M4 modelling for multilingual transfer learning. This technique, as Mayhew et al., (2017, p. 3) called it, is a "cheap translation" method to get more data (even more than back-translation itself), as no monolingual data in Spanish was back-translated to create an English system, nor used parallel data. This synthetic new dataset will be used to train NER systems, employing different methods. NMT systems, especially those that are not domain-specific, i.e., Google, are prone to errors, so this

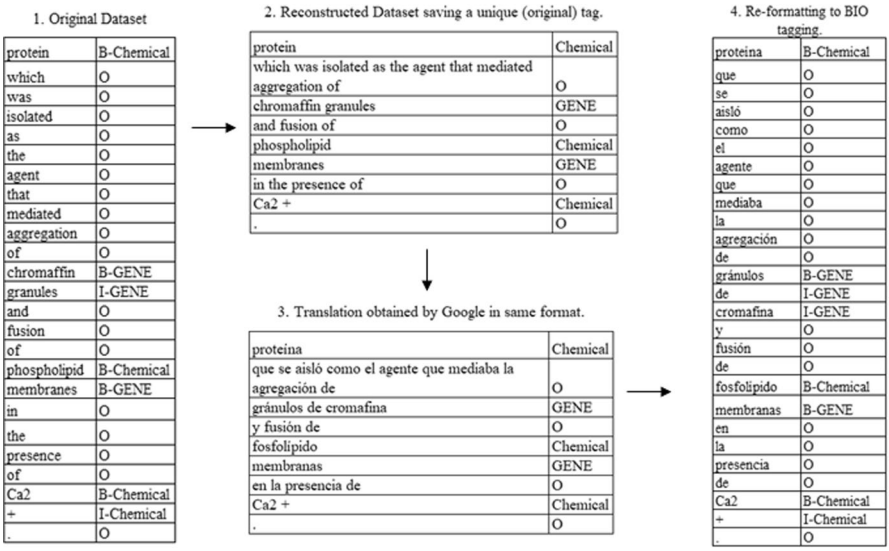

Fig. 1 Steps involved in the translation process of a real example taken from the CRAFT corpus

new MT translated dataset is not intended to be a gold-standard corpus for developing systems for Spanish only.

Mayhew et al. (ibid.) used dictionaries and Google to translate an English dataset. Their dataset was constructed by translating sentences one at a time, using fast_align to obtain alignments, and then projecting the tags, which, according to the authors, resulted in errors in the tag projection and noise. We used a similar process to create a new dataset for Spanish without the use of dictionaries. Because the CRAFT dataset is preformatted in the CONLL format, we reconstructed all of the words with an "O" tag to make real sentences, as well as multi-word NEs. We fed Google Translate with the sentences separated from the NEs, either single words or multi-words. Then, we mapped back the tags to avoid errors in the misalignment of sentences. The translation process of this dataset is described as follows:

- Concatenate words with the same tag to create sentences, as NMT systems tend to create better output when a sentence is provided instead of a single word.
- Preserve a unique tag per sentence for future mapping. Step 2 in Fig. 1 shows that by grouping tokens with the same tag and recreating sentences, a single tag is kept; removing prefix "B" or "I" from tag, and so "chromaffin granules" tagged as [chromaffin B-Gene, granules I-Gene] will have the only Gene Tag. As translation might change the order of the words, the beginning of the entity might change and so the "B" prefix. In this example, the Spanish translation "gránulos de cromafina" will later be reconstructed by assigning the beginning tag to "gránulos" instead of "cromafina" as in the English sentence.
- Pass the datasets (training, test, and development) onto Google translator in an.xlsx format to preserve the order of the sentences and assure the tags match.

Table 1 Comparison of an original sentence from the CRAFT corpus and the output of the same sentence after using a commercial MT system

English		Spanish	
Token	Tag	Token	Tag
A	O	Un	O
three	O	modelo	O
-	O	tridimensional	O
dimensional	O	de	O
model	O	CLN2	B-Protein
of	O	humano	I-Protein
human	B-Protein	fue	O
CLN2	I-Protein	construido	O
was	O	en	O
built	O	base	O
based	O	principalmente	O
mainly	O	a	O
on	O	la	O
the	O	homología	B-Sequence
homology	B-Sequence	con	O
with	O	Pseudomonas	B-Taxon
Pseudomonas	B-Taxon	sp	O
sp	O		O
	O		

– Collect the output from Google Translate.
– Assign the labels matching the original annotation previously stored. Since the reconstruction created sentences, a single tag is assigned per sentence.
– Tokenise each sentence and assign the proper tag using the BIO scheme. All "O" sentences keep only the "O" tag, whereas the entities get "B" for the beginning of the entity and "I" for the rest of the tokens in composed entities.

Figure 1 shows the translation process of a real sentence from the development subset of the CRAFT corpus.

The process creates a new translated Spanish CRAFT dataset. Polyglot systems tend to preserve essential weights for each language, as well as sharing parameters, according to Mueller et al. (ibid.). As a result, the construction of this dataset in Spanish is a strategy to improve performance in a transfer learning environment. Table 1 shows an example of the final output of a sentence in the original English dataset and the output obtained from the translation system.

Due to the sentence grouping and reconstruction process, we anticipate some errors in translation. As the sentences are split before and after an entity, some errors in gender, number, and order might arise. The emphasis is placed on the accurate reconstruction of multi-word entities and the precise alignment of the original dataset with the synthetic dataset to ensure confidence in the tag projection from English to Spanish. Although the translation technique is not fully automated, it guarantees

that the noise caused by aligners and projecting tags is completely avoided, resulting in a significantly more reliable synthetic dataset. Further testing will try to mitigate such noise by employing pre-trained models and transfer learning.

3.1.2 Entity replacement

In order to increase system performance and create a robust cross-lingual NER system, we follow Liu et al.'s (2020) approach for data augmentation: replacement of entities. Random replacement of existing entities with unseen entities is used to create new datasets for a system. We compiled a set of entities based on the official ontologies that were used to annotate the entities in the original CRAFT corpus. To ensure that all of the objects retrieved from ontologies are, in fact, new, this list has been cross-checked against the original dataset's vocabulary. The results for each tag are as follows:

- Protein[2]: 2000 entities
- Chemicals[3]: 1200 entities
- Taxonomy[4]: 1000 entities
- Gene[5]: 1000 entities
- Cell[6]: 1000 entities
- Sequence[7]: 2646 entities.

There are 8846 entities in total. The data gathered from the official ontologies is open-source and available for download on their websites.

To mirror the original format of the corpus, the list of monolingual English entities is formatted into a CONLL format with BIO scheme: [ENTITY, TAG]. 20% of the entities in the training, test, and evaluation sets have been replaced at random from the list corresponding to the tag. We attempted to replace the same number of entities for each tag to include a balanced number of new entities. 7,161 entities were altered in the training set, 5,941 in the test set, and 2,267 in the evaluation set. This was saved as a new dataset to be used in current experiments.

We concatenated some of the datasets to make different training instances that may be used to measure continuous training and compare it to dataset concatenation.

We have a total of five datasets at this point:

1.CRAFT English (Original).
2.CRAFT Spanish (MT translated).
3.CRAFT EN + ES (concatenated).
4.CRAFT EN + EN Augmented (concatenated).

[2] https://proconsortium.org/.
[3] https://www.ebi.ac.uk/chebi/downloadsForward.do.
[4] https://www.ncbi.nlm.nih.gov/taxonomy.
[5] http://geneontology.org/docs/go-archives/.
[6] https://bioportal.bioontology.org/ontologies/CL.
[7] http://www.sequenceontology.org/.

5.CRAFT EN + ES + EN Augmented (concatenated).

The CRAFT EN + ES is a dataset that combines English and Spanish data. There are 3,596 distinct entities in total. The CRAFT EN + EN Augmented specifies the concatenation of the original dataset with the augmented version obtained by entity replacement. It has 4,876 distinct entities. The final and largest dataset, CRAFT EN + ES + EN Augmented, is a concatenation of the original English dataset, the Spanish version, and the augmented English version. It has a total of 6,131 distinct entities.

3.2 Named entity recognition systems training

Inspired by the success of employing a monolingual system to evaluate its capabilities in another language (Sun & Yang, 2019) and (Hakala & Pyysalo, 2019), we employed the pre-trained "BioBERT," a variant of the well-known BERT in English pre-trained on biomedical texts. The second transformer option is the "Roberta-base-biomedical-clinical-es" for Spanish domain-specific corpora, which is a pre-trained RoBERTa-based transformer. Even though the literature has shown promise for zero-shot transfer, we opted to investigate transfer learning approaches to develop more robust systems and a consolidated bilingual unique system. This study aims to verify whether the efficacy demonstrated in back-translation, coupled with the advantages of transfer learning, could be extended to the domain of Named Entity Recognition (NER).

The presence of noise in synthetic datasets is an inherent challenge. As indicated in prior studies, transfer learning allows systems to retain weights from both languages (and datasets), share weights between languages, and handle both languages more efficiently without having to use two separate systems (Mueller et al., ibid, p. 8101). This study will test this strategy in NER. On the one hand, we will test if noise and errors produced by the cheap translation affect the F-scores of the final NER models or if the use of pre-trained Large Language Models with transfer learning are capable of using its learnt knowledge to prevail before the noise. On the other, as one of the main issues of biomedical NER is the unknown words, we introduce a higher vocabulary through augmentation. Also, these systems can benefit from learning words from two different languages in different training instances.

We conducted two types of fine-tuning: direct fine-tuning and continuous training. Direct fine-tuning entails using concatenated data from previously constructed augmented datasets, such as the CRAFT EN + ES dataset. This fine-tuning strategy simply employs one system and one dataset to produce a single output, thus no additional training or fine-tuning is required. Continuous training, on the other hand, entails fine-tuning a model with an initial dataset, e.g. CRAFT EN, and then fine-tuning the resulting system with a new dataset, e.g. CRAFT ES. Because augmentation methods have not been employed in the biomedical NER task, to the best of our knowledge, it served as impetus to evaluate both training strategies.

Since transfer learning is prone to catastrophic forgetting, in which a portion of the original weights is replaced by the new training, we chose to test both training strategies in order to compare the performance of the different datasets in different

types of training. We have trained fourteen systems: two base systems on the original English-only dataset, six systems using direct fine-tuning with the concatenated datasets, and the remaining six systems using the continuous training approach for English plus Spanish and entity replacement with the enhanced datasets (EN augmented).

All systems were fine-tuned with the following hyperparameters: learning rate: 3e-05, train batch size: 8, optimiser: Adam, betas: 0.9, 0.99, epsilon: 1e-08, epochs: 4. The following is a list of the names that will be used to refer to them.

Base models:

1. BioBERT EN
2. RoBERTa EN

Direct fine-tuning:

1. EN + ES (concatenated)
2. EN + EN Augmented (concatenated)
3. EN + ES + EN Augmented (concatenated)

Continuous training:

1. EN + ES
2. EN Augmented + EN
3. EN Augmented + ES

For each transformer, each training approach was used once. The results of the training will be discussed in Sect. 4.

4 Results

NER training is processed using various combinations of dataset and fine-tuning approach. We will describe the two training approaches: direct fine-tuning and continuous training. The implemented systems are available on HuggingFace.[8]

Direct fine-tuning uses concatenated datasets to feed the transformer and produce all the training in the specified epochs without using one system's trained weights to initialise the next one. Six systems were trained: EN + ES, EN + EN Augmented (concatenated), and EN + ES + EN Augmented (concatenated). For each of the previous dataset combinations, one BioBERT model and one RoBERTa-ES model were trained.

Continuous training uses a pre-trained system as a starting point for fine-tuning another one, making use of its learnt weights. First, we fine-tuned the models on the

[8] https://huggingface.co/StivenLancheros

Table 2 Results on the different training methods for the NER system. It shows the continuous training models and direct fine-tuning models

	System	Precision	Recall	F1
Direct Fine-tuning	BioBERT- ES Concatenated	84.87	84.43	84.65
	Roberta-ES-Concatenated	85.59	84.25	84.91
	BioBERT -EN-Augmented (concatenated)	81.22	84.75	82.94
	Roberta-EN-Augmented (concatenated)	80.77	82.58	81.67
	BioBERT-EN + ES + EN Augmented (concatenated)	82.76	84.11	83.43
	Roberta-EN + ES + EN Augmented (concatenated)	82.98	83.05	83.01
Continuous training	BioBERT- ES Continuous	85.55	85.39	85.47
	Roberta-ES- Continuous	86.64	85.87	86.25
	BioBERT- EN Augmented + EN	86.17	85.28	85.05
	Roberta- EN Augmented + EN	83.66	85.13	84.39
	BioBERT- EN Augmented + ES	85.35	84.76	85.72
	Roberta- EN Augmented + ES	**86.66**	**86.14**	**86.39**
	BioBERT base	83.97	83.66	83.82
	Roberta-ES base	82.53	81.47	82.0

first dataset and using the weights of the trained model as initialiser, we fine-tuned it with the second dataset. The two base models at the bottom of Table 2 are the starting point for the continuous training with the different datasets. Six systems were trained: EN + ES, EN Augmented + EN, EN Augmented + ES. For each of the previous dataset combinations, one BioBERT model and one RoBERTa-ES model were trained.

Data augmentation has increased the performance of all systems; thus, it is safe to say that it is beneficial for training. The success of continuous training over direct fine-tuning can be attributed to the additional fine-tuning phase, as the direct fine-tuning approach lacks the additional training instances carried out in the continuous training strategy. It should be noted that catastrophic forgetting and adopting strategies to prevent it, such as Elastic Weight Consolidation (EWC), are outside the scope of this paper and will be pursued in the future.

Concatenating the datasets gives competitive results but not the highest possible scores. Nonetheless, the best scores using concatenation were the ones without further data augmentation (only the Spanish dataset). This suggests that using concatenation with data augmented datasets, at least with the proposed augmentation technique, does not yield satisfactory results and even results in lower performance than the base models. The basic concatenated models (EN-ES) preserved the top scores in the direct fine-tuning category. However, the fine-tuning times doubled the ones in the continuous training, which makes continuous training more suitable for local training.

The translation of the dataset to create a Spanish dataset showed to be beneficial in terms of enhancing the model's knowledge and providing higher F1 scores. It is worth pointing out that the two systems with the highest scores, 86.39 and 86.25, respectively, were trained on the Spanish dataset. Surprisingly, concatenation of the

dataset plus the augmented datasets revealed that the performance figures do not rise but rather drop when employing concatenation. Table 2 shows the results of all the training instances with the different datasets. Both base models at the bottom of the table are monolingual. The results are reported on the evaluation dataset.

Most of the existing research on Named Entity Recognition (NER) using the CRAFT dataset makes use of OGER (OntoGen's Entity Recogniser), a library/method implemented to perform dictionary-based NER given a terminology list, ontology, or dictionary. Rinaldi, Basaldella, and Furrer have conducted most of the research using the CRAFT dataset and OGER in various settings. Rinaldi et al. (2017) utilised the same ontologies as CRAFT to collect and build a dictionary and employed OGER to match the entities in the text with a neural network (NN) acting as a post-filter for the dictionary annotation. In continuation, Basaldella et al. (2017) adopted the same architecture and added a distiller to filter relevant entities before passing them to the NN. The authors pointed out that "Since the NN output depends heavily on OGER's input, many of its mistakes are caused by the quality of the dictionary matching." OGER++employs a disambiguation filter by training an NN on the CRAFT corpus and, in addition to acting as a filter, provides a probability distribution over all entity types (Furrer et al., 2019a). Furrer et al. (2019b) revisited the NER task with CRAFT by testing it with BioBERT and BiLSTM architectures. These tests aimed at predicting not only entities but also identifiers and combinations of entity predictions with annotations from OGER. The best results were obtained with BioBERT. Finally, Furrer et al. (2021) tested NER+NEN (Named Entity Normalisation) parallel training on a BioBERT architecture, using OGER as before, and tested harmonisation techniques to combine the predictions of two classifiers into one.

Crichton et al. (2017) used a Multi-task learning method, employing 15 datasets, with each dataset representing a task, including AnatEM, BC2GM, BioNLP09, and others. The goal was to take advantage of transfer learning by sharing layers of a Convolutional Neural Network (CNN) among datasets, similar to the known transfer learning capabilities of transformer architectures such as BioBERT used in our research. They reported increases in F1 scores on the multi-task models compared to the single-task (single dataset) models of 6.3% for some datasets and 1.1% for others. Lastly, the popular spaCy python library reports a NER model based on a transition system using a chunking model, and presents findings on several datasets, offering two models that, although not outperforming any models in the literature, are considered competitive baselines for 5 of the 9 datasets.

Comparable to SpaCy, our research diverges from using external resources during training and inference. Research based on OGER involves the construction of a dictionary or ontology to perform dictionary matching. While we extracted lists of entities from the corresponding ontologies to perform data augmentation, our model does not require dictionary matching and all the preprocessing steps associated with it. Leveraging the power of transformer architectures, data augmentations, and transfer learning, we aimed at training a model capable of recognising patterns of how entities present themselves in texts, not only in one language but in two.

As shown in Table 3, in comparison to the second-best-scoring system, the top system outperforms systems in the literature for the same dataset by+7.33

Table 3 F1 scores reported in the literature compared to the best performer obtained by this study. All models were evaluated with the CRAFT corpus

System	F1-score	Architecture
Rinaldi et al. (2017)	70	OGER + NN (dictionary based)
Basaldella et al. (2017)	73	NN + OGER + CRF
Crichton et al. (2017)	79.56	CNN Multitask model
Furrer et al., (2019a, 2019b)	71.4	OGER + + (Dictionary and corpus based + NN)
Neumann et al. (2019)	78.35	SciSpacy
Furrer et al., (2019a, 2019b)	79.40	BioBERT + ids + spans + OGER (run 3)
Furrer et al. (2021)	**86.85**	BioBERT + OGER
Our best performer	86.39	RoBERTa

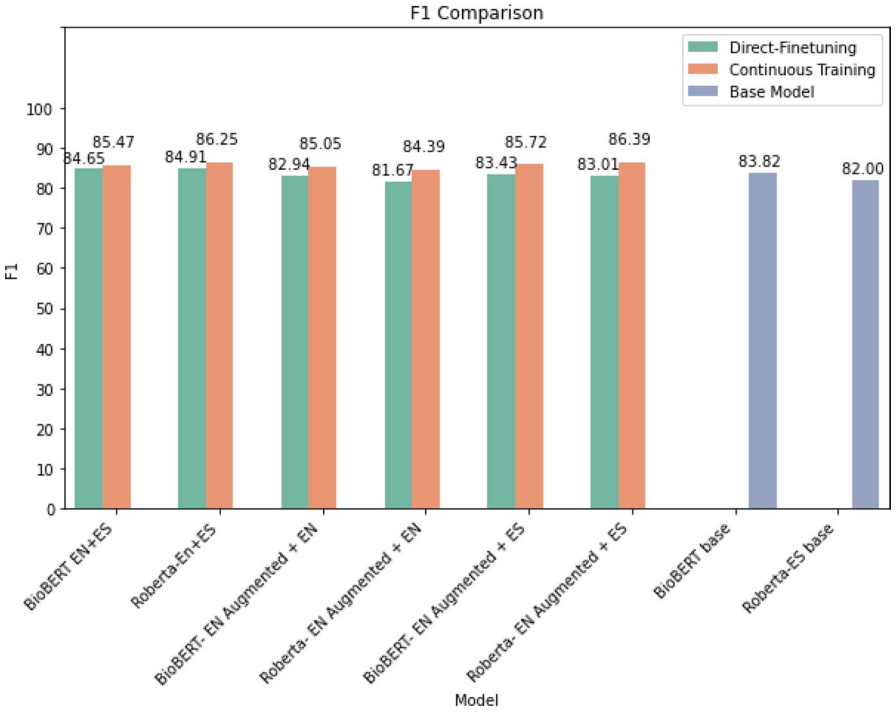

Fig. 2 F1 scores of the 14 NER systems trained on different fine-tuning methods and with different dataset combinations

F1 score points (Crichton et al., 2017). The top-performing system by Furrer et al. (2021) is a combination of BioBERT and OGER, which uses a dictionary-based approach and fares better than our best-performing system by only 0.46 F1 score points. Given that our system will perform in another language and will not employ dictionary-based techniques, these competitive results hold promise (Fig. 2).

5 Conclusions

One of the main contributions of this study is the use of an English dataset in the biomedical domain to create a bilingual system. This NER system not only identifies NEs in both languages, but it does so using one of the datasets with the highest number of NEs (6 classes), where most datasets only contain a small number of NE classes based on the task at hand, such as diseases and chemicals (BC5CDR) or genes and proteins (GENIA). Even fewer NE classes are available in Spanish datasets, such as Chemical and proteins (PharmaCoNER) and Cancer concepts (CANTEMIST), both of which have a significantly different annotation scheme. Furthermore, NEs contained in CRAFT have not been annotated in Spanish texts. The synthetic dataset for Spanish was created using a "cheap translation" inspired by the success of back translation for NMT. The F1 score difference from the best performer, as shown in Table 3, is just 0.46 F1-score points, surpassing most systems in the literature and placing second, proving that the NER system's training was successful. In contrast to the best performance reported in literature, our system is able to recognise NEs in two languages, EN and ES, and it is independent of dictionaries or external information.

Entity replacement was used as a second augmentation technique. Entities of the same tag were replaced with data extracted from the official ontologies employed by the annotators of CRAFT, as one of the issues was the OOV words, which are common in the biomedical domain due to the constant generation of terms. As a result, a different CRAFT corpus was generated with 20% of its entities replaced. In the future, alternative percentages of replacement might be evaluated and reported, as well as an attempt to replace the entire dataset.

In total, fourteen systems were trained for NER using a variety of dataset combinations, including the original dataset plus the Spanish version or the new CRAFT in English. We employed either direct fine-tuning by concatenating the datasets or continuous fine-tuning by sequentially training the separate datasets and using the transfer learning success of transformers. Systems trained via transfer learning performed best, while systems trained by concatenating datasets showed a downward trend in F-score. As in Lamurias & Couto (2019), the performance of the system did improve by combining datasets, but the transfer learning technique proved to be better. The scarcity and difficulty of biomedical annotated data in other languages can be solved by combining data augmentation with transfer learning, which also uses state-of-the-art systems. Such a method might be applied to different languages to improve results, using in-domain data without relying entirely on the transformers' zero-shot capabilities, as well as for languages where zero-shot is not an option.

Our novel methodology will make it possible for another language to benefit from one of the biggest biomedical datasets for NER. This dataset, and methodology, can be beneficial for researchers and future studies on other datasets to cover more NE classes. Translations of the dataset in other languages can be used, either from English to the target language or by leveraging the already translated Spanish dataset to obtain another in a close (Romance) language. The proposed method in this research is applicable to larger Language Model architectures (LLMs) and is compatible with

current quantization and training enhancement techniques. This approach is portable to other languages, provided any modern MT system has support for it. Incorporating an EWC into the training process to prevent catastrophic forgetting and increasing the learning of new weights in the NER system is another potential future project.

Acknowledgements The present research has been partially carried out in the framework of the VIP II project (PID2020-112818GB-I00).

Author contributions BSL designed and performed the experiments, derived the models, analysed the data, and wrote the main manuscript. Both GCP and RM contributed with original ideas and input to methodology, worked on the final version of the manuscript and supervised the project. GCP also contributed to the experimental design, with special reference to translation issues. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Data availability The datasets generated during the current project are available in the Open Science Framework repository: https://github.com/stivennpe/CrossLingual-BioNER

Declarations

Competing interests The authors declare no competing interests.

References

Appelt, D., Hobbs, J., Bear, J., Israel, D., Kameyama, M., Martin, M., Myers, K., & Tyson, M. (1995). SRI International FASTUS system: MUC-6 test results and analysis. In *Proceedings of the 6th conference on message understanding (MUC6 '95).* Association for Computational Linguistics . USA, pp. 237–248. https://doi.org/10.3115/1072399.1072420

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Jr., Cohen, K. B., Verspoor, K., Blake, J. A., & Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics, 2012*(13), 161. https://doi.org/10.1186/1471-2105-13-161

Basaldella, M., Furrer, L., Tasso, C., & Rinaldi, F. (2017). Entity recognition in the biomedical domain using a hybrid approach. *Journal of Biomedical Semantics, 8*(1), 51. https://doi.org/10.1186/s13326-017-0157-6

Beltagy , I., Lo, K., & Cohan, A. (2019). SCIBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* . pp. 3615–3620, Hong Kong, China, November 3–7, 2019. https://aclanthology.org/D19-1371.pdf

Carrino, C. P. , Armengol-Estapé, J., Gutiérrez-Fandiño, A., Llop-Palao, J., Pàmies, M., Gonzalez-Agirre, A., & Villegas, M. (2021). Biomedical and clinical language models for Spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. arXiv:2109.03570.

Cho, H., & Lee, H. (2019). Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics, 20*, 735. https://doi.org/10.1186/s12859-019-3321-4

Crichton, G., Pyysalo, S., Chiu, B., & Korhonen, A. (2017). A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics, 18*(1), 368. https://doi.org/10.1186/s12859-017-1776-8

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics, 47*, 1–10. https://doi.org/10.1016/j.jbi.2013.12.006

Edunov, S., Ott, M., Auli, M., & Grangier. D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. pp. 489–500, Brussels, Belgium. Association for Computational Linguistics. https://aclanthology.org/D18-1045/

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D., & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence 165, pp. 91–134, Essex: Elsevier Science Publishers. https://homes.cs.washington.edu/~etzioni/papers/knowitall-aij.pdf

Furrer, L., Jancso, A., Colic, N., & Rinaldi, F. (2019a). OGER++: Hybrid multi-type entity recognition. *Journal of Cheminformatics, 11*(1), 7. https://doi.org/10.1186/s13321-018-0326-3

Furrer, L, Cornelius, J., & Rinaldi, F. (2019b). UZH@CRAFT-ST: A sequence-labeling approach to concept recognition. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pp. 185–195, Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-5726/?msclkid=c75ea2e9b1be11ec9a776f11371b2847.

Furrer, L., Cornelius, J., & Rinaldi, F. (2021). Parallel sequence tagging for concept recognition. *BMC Bioinformatics, 22*(1), 1–18.

Grishman, R. (1995). The NYU system for MUC-6 or where's the syntax? In *Proceedings of the 6th conference on Message understanding (MUC6 '95)*. Association for Computational Linguistics, USA, pp. 167–175. https://doi.org/10.3115/1072399.1072415

Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., & Martin Krallinger. (2019). PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pp. 1–10, Hong Kong, China. https://aclanthology.org/D19-5701/

Habib, M., & van Keulen, M. (2012). Unsupervised improvement of named entity extraction in short informal context using disambiguation clues. Faculty of EEMCS, University of Twente, Enschede. The Netherlands. http://ceurws.org/Vol-925/paper_1.pdf

Hakala, K., & Pyysalo, S. (2019). Biomedical named entity recognition with multilingual BERT. In *Proceedings of the 5th workshop on BioNLP open shared tasks* . pp. 56–61, Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-5709

Hira, H., Rauf, S., Kiani, K., Zafar, A., & Nawaz, R. (2019). Exploring transfer learning and domain data selection for the biomedical translation. In *Proceedings of the fourth conference on machine translation (Volume 3: Shared Task Papers, Day 2)*, pp. 156–163, Florence, Italy. Association for Computational Linguistics. https://aclanthology.org/W19-5419/

Jofche, N., Mishev, K., Stojanov, R., Jovanovik, M., Zdravevski, E., & Trajanov, D. (2022). Named entity recognition and knowledge. Extraction from pharmaceutical texts using transfer learning. *Procedia Computer Science., 203*, 721–726. https://doi.org/10.1016/j.procs.2022.07.107

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., Sayle, R. A., Batista-Navarro, R. T., Rak, R., Huber, T., Rocktäschel, T., Matos, S., Campos, D., Tang, B., Xu, H., … Valencia, A. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics, 7*(1), 2. https://doi.org/10.1186/1758-2946-7-S1-S2

Lamurias, A., & Couto, F. (2019). LasigeBioTM at MEDIQA 2019: Biomedical question answering using bidirectional transformers and named entity recognition. In *Proceedings of the 18th BioNLP workshop and shared task, pp. 523–527, Florence, Italy*. Association for Computational Linguistics . https://aclanthology.org/W19-5057

Leaman, R., & Lu, Z. (2016). TaggerOne: Joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics, 32*(18), 2839–2846. https://doi.org/10.1093/bioinformatics/btw343

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., & Lu, Z. (2016). BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database: the Journal of Biological Databases and Curation*. https://doi.org/10.1093/database/baw068

Li, L., Zhao, J., Hou, L., Zhai, Y., Shi, J., & Cui, F. (2019). An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. *BMC Medical Informatics and Decision Making, 19*, 235. https://doi.org/10.1186/s12911-019-0933-6

Li, P., Wang, M., & Wang, J. (2020). Named entity translation method based on machine translation lexicon. *Neural Computing and Applications, 2021*(33), 3977–3985. https://doi.org/10.1007/s00521-020-05509-y

Liu, Q., Li, P., Lu, W., & Cheng., Q. (2020). Long-tail dataset entity recognition based on Data Augmentation. In *EEKE 2020 - workshop on extraction and evaluation of knowledge entities from scientific documents*. http://ceur-ws.org/Vol-2658/paper10.pdf

Mayhew, S., Tsai, C., & Roth, D. (2017). Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics. https://aclanthology.org/D17-1269

Microsoft. (n.d.). Biomedical NLP group. https://www.microsoft.com/en-us/research/group/biomedical-nlp-group/?msclkid=f6698204ba9511ecb36d9abbaf1a0a2e

Miranda-Escalada, A., Farré-Maduell, E., and Krallinger, M. (2020). Named entity recognition, concept normalization and clinical coding: Overview of the Cantemist track for cancer text mining in spanish, corpus, guidelines, methods, and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR workshop proceedings*, pp. 303–323 (2020). http://ceur-ws.org/Vol-2664/cantemist_overview.pdf

Mueller, D., Andrews, N., & Dredze, M. (2020). Sources of transfer in multilingual named entity recognition. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. pp. 8093–8104 July 5–10, 2020. https://aclanthology.org/2020.acl-main.720.pdf

Nadeau, D., Turney, P., & Matwin, S. (2006). Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity. In *Proceedings of the 19th international conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence (AI'06)*. Springer, Berlin. pp. 266–277. https://doi.org/10.1007/11766247_23.

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP workshop and shared task* . PP. 319–327, Florence, Italy. Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-5034

Ponomareva, N., Rosso, P., Pla, F., & Molina, A. (2007). Conditional random fields vs. hidden M&arkov models in a biomedical named entity recognition task. In *Proceedings of the RANLP'07 conference*. Bulgaria, Borovets. 2007, 3.

Rinaldi, F., Furrer, L., & Basaldella, M. (2017). Efficient and accurate entity recognition for biomedical text. In: *BioCreative VI workshop*, Bethesda, MD, USA, 18 October 2017–20 October 2017, pp. 195–197. https://www.zora.uzh.ch/id/eprint/141515/?msclkid=8bfd938ab1b611ec8a1140cdaceb46c5

Sahu, S., & Anand, A. (2016). Recurrent neural network models for disease name recognition using domain invariant features. In *Proceedings of the 54th annual meeting of the association for computational linguistics* . pp. 2216–2225, Berlin, Germany, August 7–12, 2016. https://aclanthology.org/P16-1209.pdf

Sun, C., & Yang, Z. (2019). Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pp. 100–104 Hong Kong, China, November 4, 2019. Association for Computational Linguistics. https://aclanthology.org/D19-5715.pdf

Saunders, D., Stahlberg, F., and Byrne, B. 2019. UCAM Biomedical Translation at WMT19: Transfer Learning Multi-domain Ensembles. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pp. 169–174, Florence, Italy. Association for Computational Linguistics.

Tanabe, L., Xie, N., Thom, L. H., Matten, W., & Wilbur, W. J. (2005). GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics, 6*(1), S3. https://doi.org/10.1186/1471-2105-6-S1-S3

Wang, X., Zhang, Y., Ren, X., Zhang, Y., Zitnik, M., Shang, J., Langlotz, C., & Han, J. (2019). Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics (oxford, England), 35*(10), 1745–1752. https://doi.org/10.1093/bioinformatics/bty869

Weischedel, R. (1995). BBN: Description of the PLUM System as Used for MUC-6. Proceedings of the Sixth Message Understanding Conference (MUC-6), pp. 55–69, Columbia, Maryland: Morgan Kaufmann Publishers, Inc. https://www.aclweb.org/anthology/M95-1006.pdf

Zhang, L., Pan, Y., & Zhang, T. (2004). IBM Research Report Focused Named Entity Recognition Using Machine Learning. RC23066 (C0401-004) January 19, 2004, Computer Science . https://domino.oweb.draco.res.ibm.com/reports/rc23066.pdf

Zhao, S., Su, C., Lu, Z., & Wang, F. (2021). Recent advances in biomedical literature mining. *Briefings in Bioinformatics, 22*(3), 057. https://doi.org/10.1093/bib/bbaa057

Authors and Affiliations

Brayan Stiven Lancheros[1] · Gloria Corpas Pastor[2] · Ruslan Mitkov[3]

✉ Brayan Stiven Lancheros
b.s.lancherosrincon@wlv.ac.uk

✉ Ruslan Mitkov
r.mitkov@lancaster.ac.uk

Gloria Corpas Pastor
gcorpas@uma.es

[1] University of Wolverhampton, Wolverhampton, UK

[2] Universidad de Malaga, IUITLM, Malaga, Spain

[3] Lancaster University, Lancaster, UK