Contributions to Clinical Named Entity Recognition in Portuguese

Fábio Lopes
CISUC, DEI
University of Coimbra
Portugal
fadcl@student.dei.uc.pt

César Teixeira
CISUC, DEI
University of Coimbra
Portugal
cteixei@dei.uc.pt

Hugo Gonçalo Oliveira
CISUC, DEI
University of Coimbra
Portugal
hroliv@dei.uc.pt

Abstract

Having in mind that different languages might present different challenges, this paper presents the following contributions to the area of Information Extraction from clinical text, targeting the Portuguese language: a collection of 281 clinical texts in this language, with manually-annotated named entities; word embeddings trained in a larger collection of similar texts; results of using BiLSTM-CRF neural networks for named entity recognition on the annotated collection, including a comparison of using in-domain or out-of-domain word embeddings in this task. Although learned with much less data, performance is higher when using in-domain embeddings. When tested in 20 independent clinical texts, this model achieved better results than a model using larger out-of-domain embeddings.

1 Introduction

In recent years, much data has been produced on different areas, including healthcare, which, besides its general relation to well-being, is also economically-relevant (Folland et al., 2017). We focus on the clinical field, where valuable information is hidden on produced admission notes, diagnostic test reports, patient discharge letters or clinical case reports. The latter contain information about patient clinical histories, such as their condition; diagnostic tests and respective results; or treatments and how they were administered. Such data is very useful for clinical professionals in their future decisions about what diagnostic tests or therapies a patient has to do, based on past clinical information. However, manually processing all available texts and looking for important information is impractical for humans. To make it more tractable, Natural Language Processing (NLP) tools have been developed for automating tasks such as Information Extraction (IE), including Named Entity Recognition (NER), and ultimately store acquired information in relational databases, where queries should be more efficient.

Similarly to many other NLP-related tasks, the field of clinical NLP has been growing. This is both reflected in the organization of shared tasks (Uzuner et al., 2011; Stubbs and Uzuner, 2015; Doğan et al., 2014; Pestian et al., 2007; Elhadad et al., 2015; Bethard et al., 2016; Kelly et al., 2016), which made available several datasets, such as Informatics for Integrating Biology & the Bedside (i2b2); or in the adoption of deep neural network architectures that lead to state-of-the-art results, namely Bidirectional Long Short Term Memory with a stacked Conditional Random Fields layer (BiLSTM-CRF) (Xu et al., 2017; Unanue et al., 2017). However, most of the work going on targets text written in English. When it comes to other languages, such as Portuguese, the number of studies on this field is much lower (Névéol et al., 2018).

This work aims to boost clinical NLP in Portuguese with three main contributions: (i) A collection of Portuguese clinical texts with manually-labelled named entities; (ii) A model of word embeddings learned from a larger collection of Portuguese clinical text (i.e., Neurology clinical case descriptions); (iii) An analysis of the performance of state-of-the-art models in Portuguese clinical NER, namely BiLSTM-CRF neural networks (Lample et al., 2016), tested on the labelled collection, either using the previous word embeddings or general-language word embeddings.

In the next section, we introduce deep learning architectures and word embedding (WE) models that have been used in NER. Section 3 describes how texts were labelled and provides some figures on the resulting dataset and its revision. Furthermore, we explain how the in-domain WE model was trained and its qualitative difference towards

Proceedings of the BioNLP 2019 workshop, pages 223–233
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

the pre-trained out-of-domain WE model used. Finally, we explain the architecture of our deep learning model. Section 4 reports the results for hyperparameters grid search. After choosing the best model for both in-domain and out-of-domain WEs, we tested it on an independent test set. We report micro-averaged relaxed F1-score and strict F1-score of 70.41% and 62.71%, respectively. We conclude with a brief discussion.

2 Related Work

Training a model for clinical NER requires access to much clinical textual data. Although much text of this kind is produced everyday, its availability is highly limited due to strict ethical regulations that constrain using data with personal information, as in clinical case or diagnostic test reports. Still, when available, such texts constitute valuable sources of data, and may be used in the development of models for Information Extraction, including Named Entity Recognition (NER).

In order to create machine learning models that identify and classify named entities (NEs), the latter have to be annotated on a collection of texts, which can be used as training and/or testing data. That is generally done manually, as several authors did. For instance, Uzuner et al. (2011) annotated 871 medical records with Medical Problems, Treatments and Tests, in order to provide a dataset for the 2010 i2b2/VA concept extraction shared task; and Stubbs and Uzuner (2015) labelled 1,304 individual longitudinal records with heart-risk NEs (e.g. Diabetes references or Hypertension) with 0.95 agreement ratio. Beyond English, some studies involved the creation of datasets in other languages. Skeppstedt et al. (2014) annotated Disorders, Findings, Body Structures and Pharmaceutical Drugs, in 1,104 clinical notes in Swedish, with agreement ratios of 0.79, 0.66, 0.80 and 0.90, respectively. Mykowiecka et al. (2009) annotated 700 mammography reports and 100 diabetic discharge documents, in Polish, with NEs that carry information about Pathological Findings, Breast Tissue, and Crucial Health information about diabetic patients. Ferreira et al. (2010) manually labelled 90 clinical notes in Portuguese with NEs such as Condition, Anatomical Site and Finding. Although in Portuguese, the previous dataset is not publicly available due to ethical regulations, but the annotation guidelines followed are published (Ferreira, 2011).

In recent years, deep learning approaches have been used for NER, leading to state-of-the-art results. Clinical NER is not an exception, with such models used for extracting data from Electronic Medical Records (EMR). Adopted architectures include Recurrent Neural Networks (RNN), with simple RNN layers, LSTM layers, BiLSTM layers or Gated Recurrent Unit (GRU) layers; Convolutional Neural Networks (CNN); and also Feed-Forward Networks (FFN). Luu et al. (2018) showed that a vanilla RNN outperforms a FNN using the same features on clinical texts provided in the CLEF eHealth 2016 task (Kelly et al., 2016) on the extraction of relevant information from nursing shift changes notes. This was expected because FNNs do not consider past information.

Chokwijitkul et al. (2018) assessed the performance of CNN, RNN, LSTM, BiLSTM and GRU networks for identifying heart risk factors in EMRs and found that BiLSTM networks achieved the best F-measure. They further show that such models perform near the rule-based and shallow machine learning models, but do not resort to gazetteers or knowledge bases. Wu et al. (2018) compared different classifiers (CRF, CNN and BiLSTM) for NER, using the dataset of the 2010 i2b2 NLP challenge. They also compared their models with the best models at the time (Structured SVM) and trained during the competition (Semi-Markov model), and used pre-trained word embeddings (WEs) as features for the BiLSTM network and the CNN. For the CRF, they used three different feature sets: only word and n-gram features; the previous plus linguistic features and document level features, such as section names; and all the previous plus features from general clinical NLP systems (MedLEE, MetaMap, KnowledgeMap) and gazetteer features from the UMLS terminology. Similarly to Chokwijitkul et al. (2018), they report that the BiLSTM network outperformed all the others.

Others developed a BiLSTM network with a character embedding layer, a WE layer and a CRF layer. Xu et al. (2017) evaluated their architecture on the NCBI Disease Corpus (793 PubMed medical literature abstracts), while Unanue et al. (2017) evaluated their models with three different datasets (2010 i2b2/VA dataset, DrugBank and MedLine). Both showed that the CRF layer and the character embedding feature have great importance on the performance of a BiLSTM network.

Although these models became the trend in NER, they rely heavily on the quality of the WE models for converting each word to its embedding vector. On the clinical domain, Newman-Griffis and Zirikly (2018) compared WEs using in-domain and out-of-domain corpora. In-domain corpora consisted of two different datasets, one with 154,967 Electronic Health Records (EHR) and a subset with 17,952 EHR documents focused on Physical Therapy (PT) and Occupational Therapy (OT). Out-of-domain corpora were constituted by 14.7 million abstracts from the 2016 PubMed baseline and two million free-text documents released as part of the MIMIC-III critical care DB. Besides those, they used a FastText model, pre-trained on Wikipedia 2017 documents. They reported that, with WEs trained with small in-domain corpora, results were similar to those achieved with the large out-of-domain corpora. Unanue et al. (2017) additionally showed that re-training WE models with domain-specific texts improves the performance of the model.

Although not on the clinical domain, there is some related work on Portuguese. On general NER, de Castro et al. (2018) recently achieved state-of-art results using a BiLSTM-CRF model. On distributional similarity, Hartmann et al. (2017) compared Portuguese word WEs, learned with different methods, in both intrinsic (syntactic and semantic analogies) and extrinsic (PoS tagging and sentence similarity) tasks. There are also studies suggesting that, in tasks such as PoS tagging and NER, combining character embedding with pre-trained WE outperforms approaches that use only WEs (Santos and Zadrozny, 2014; dos Santos and Guimarães, 2015).

3 Experimental Set-up

This section presents the textual data used, the guidelines followed for its annotation and characterizes the resulting dataset with some numbers on its contents and revision. It further explains how the WE models used were learned and the architecture of the NER model, including how its hyperparameters grid search was made.

3.1 Dataset

Three different datasets were used in different stages of this work:

- For training and validation, 281 clinical case texts collected from the numbers 1 and 2 of volume 17 of the clinical journal Sinapse (Sinapse, 2017a,b), published by the Portuguese Society of Neurology. Neurology texts were used because the testing texts, that originally motivated this work, were obtained from the Neurology service.

- For testing, a small set of 20 clinical texts obtained from the Neurology service of the Coimbra University Hospital Centre (CHUC), in Coimbra, Portugal. These include admission notes, diagnostic test reports and patient discharge letters and were originally used in the development of the European Epilepsy Database (Klatt et al., 2012).

- For training the in-domain WE model, a total of 3,377 clinical texts were collected from all the volumes of the Sinapse journal, published between 2001 and 2018[1]. Although the journal contains clinical cases and experimental reports we just collected the clinical cases.

As all the texts used for training, validation and test were in a raw format, they were preprocessed with tools in NLPPort (Rodrigues et al., 2018), a NLP toolkit for Portuguese, based on OpenNLP – each text was tokenized with TokPort, PoS-tagged with TagPort, and lemmas for each token-PoS pair obtained with LemPort. After preprocessing, manual NE annotation was based on the guidelines described in Ferreira's PhD Thesis (Ferreira, 2011), originally developed with the help of physicians and linguists and used in the annotation of Ferreira's dataset. All the NEs in the guidelines were considered, with the exception of Location, because it represents geographical locations, e.g, "Coimbra" (a city) or "domicílio" (home, in Portuguese), which does not represent important clinical information. Although DateTime does not represent clinical information as well, it is important to know what temporal information is related to diseases or therapies, e.g., their frequency or duration. Furthermore, two new NE classes were introduced, namely Genetics and Additional Observations. The former was used for information about genes related to diseases (e.g., "...o estudo do *gene PMP22* identificou..." (*...study of the gene PMP22 identified...*)), and the latter for all clinically-relevant information that did not suit any of the other classes (e.g. "...medicada e

[1] http://www.sinapse.pt/archive.php

ex-fumador, refere...” (*...medicated and ex-smoker, states...*). The dataset thus considers 14 different tags, one for each NE class, plus the Out tag, for tokens not belonging to a NE. For annotation, we adopted the Inside-Outside-Beginning (IOB) format, which allows to distinguish between tokens in the beginning and inside a NE. This is essential to sequential classifiers and allows for better rules, which do not enable to tag a token as inside-NE before the beginning of the same NE. Table 1 illustrates the annotated data.

Tables 2 and 4 provide a quantitative analysis of the training and validation datasets, while tables 3 and 5 a quantitative analysis of the independent test set. Tables 2 and 3 quantify the tokens for each IOB tag (NT), the number of distinct tokens (NDT), and their ratios (NTR, NDTR). Finally, tables 4 and 5 show the number of NE occurrences (O), the number of distinct NE occurrences (DO) and their ratios (OR, DOR). As the test set has only reports related to epilepsy, it does not have occurrences of the Genetics NE.

The entire dataset was annotated by the first author of this paper, a last-year student of the MSc in Biomedical Engineering. After that, to validate the annotation, 30% of the dataset was revised by two MSc students in Biomedical Engineering, two PhD students in Data Science, one Computer Science Professor working on NLP and NER, and one Physiotherapist. Each of the previous revised 15 texts. Based on the revised subset, we calculated the agreement ratios as the ratio between the number of tokens which were annotated with the same tag as our annotation and the total number of tokens for each NE. Although there were some tokens annotated with different tags, we did not change dataset labels. Agreement ratios (ARs) for each NE, as well as the number of agreed (AT) and of not-agreed tags (NAT) are in table 6.

The lowest ARs are for Additional Observations, Characterization and Results. They were also the classes whose original labelling raised more doubts. Additional Observations is a general class which may include other NEs, for instance, in case it does not relate to the patient but to their family — e.g., “...diagnóstico de doença neoplástica no marido...” (*...diagnosis of neoplastic disease in her husband...*) — , or information about the patient that is important but does not suit any other class — e.g. “...abandono do acompanhamento médico...” (*...abandonment of medical assistance...*). Characterization may have tokens from the Condition or Evolution classes, depending on the perspective of the reader — e.g., “possível” (*possible*) in “possível processo vascular” (*possible vascular process*) or “hipótese” (*hypothesis*) in “hipótese de metástase” (*hypothesis of metastasis*), for Condition, and “progressivo” (*progressive*) in “declínio cognitivo progressivo” (*progressive cognitive decline*) for Evolution. Depending on their interpretation, results may also have tokens from Condition — e.g. “nova lesão” (*new injury*) in “...RM-CE que documentou nova lesão...” (*...RM-CE which documents a new injury...*), or “hematoma” in “...TAC-CE que mostrou aumento do hematoma...” (*...TAC-CE which shown an increase of the hematoma...*). Overall, the agreement for all the NE classes is above 90%, except for Characterization. This is high, especially considering the number of classes covered and that the used documents are not always easy to interpret, due to the high presence of medical terminology. We recall that these numbers apply for only 30% of the dataset. Due to lack of time, the remaining documents were not revised.

Token	POS Tag	Lemma	IOB Tag
de	prp	de	O
66	num	66	O
anos	n	ano	O
,	punc	,	O
com	prp	com	O
antecedentes	n	antecedente	B-DT
de	prp	de	O
dislipidemia	n	dislipidemia	B-C
e	conj-c	e	O
síndrome	n	síndrome	B-C
depressiva	adj	depressivo	I-C
,	punc	,	O
começou	v-fin	começar	O
por	prp	por	O

Table 1: Example of dataset annotation. Sentence: “...de 66 anos, com antecedentes de dislipidemia e síndrome depressiva, começou por...”

3.2 Word Embeddings

In-domain WE models were trained with 3,377 clinical texts collected from the Sinapse journal, comprising 686,762 tokens all together. For training the model, we used the FastText algorithm (Bojanowski et al., 2017), available in the Gensim library (Rehurek and Sojka, 2010). FastText learns embeddings for characters and represents each word by the sum of its characters. It was used instead of word2vec (Mikolov et al., 2013b) because, while word2vec would consider unseen

IOB Tags	NT	NTR (%)	NDT	NDTR (%)	Examples	Examples (English)
B-AS	2,491	4.272	770	6.794	seio (B-AS)	venous
I-AS	2,510	4.305	599	5.285	venoso (I-AS)	sinous
B-C	3,884	6.662	1,074	9.476	paramnésia (B-C)	reduplicative
I-C	3,634	6.233	1,269	11.196	reduplicativa (I-C)	paramnesia
B-CH	1,043	1.789	503	4.438	mais (B-CH)	more
I-CH	576	0.988	358	3.159	marcado (I-CH)	marked
B-DT	1,516	2.600	280	2.470	18 (B-DT)	18
I-DT	2,495	4.279	378	3.335	semanas (I-DT)	weeks
B-EV	794	1.362	184	1.623	desenvolveu (B-EV)	gradually
I-EV	452	0.775	120	1.059	gradualmente (I-EV)	developed
B-G	61	0.105	15	0.132	gene (B-G)	EGFR
I-G	62	0.106	47	0.415	EGFR (I-G)	gene
B-N	768	1.317	46	0.406	não (B-N)	not
I-N	2	0.003	2	0.018	impedindo (I-N)	hindering
B-OBS	217	0.372	153	1.350	restantes (B-OBS)	remaining
I-OBS	227	0.389	144	1.271	irmãos (I-OBS)	siblings
B-R	1,767	3.031	589	5.197	VS (B-R)	increased
I-R	2,520	4.322	922	8.135	aumentada (I-R)	ESR
B-RA	71	0.122	14	0.124	intravenoso (B-RA)	intravenous
I-RA	0	0.000	0	0.000		
B-T	2,041	3.501	490	4.323	estudo (B-T)	cytogenetic
I-T	2,113	3.624	677	5.973	citogénico (I-T)	study
B-THER	894	1.533	384	3.388	correção (B-THER)	correction
I-THER	709	1.216	332	2.929	de (I-THER)	of
B-V	410	0.703	276	2.435	0.8 (B-V)	0.8
I-V	584	1.002	112	0.988	células (I-V)	cells
O	26,463	45.388	1,596	14.082	-	-
Total	58,304	100,000	11,334	100.000	-	-

Table 2: Quantitative analysis of the training/validation dataset.

Reference: CH: Characterization; T: Test; EV: Evolution; G: Genetics; AS: Anatomical Site; N: Negation; OBS: Additional Observations; C: Condition; R: Results; DT: DateTime; THER: Therapeutics; V: Value; RA: Route of Administration; O: Out

IOB Tag	NT	NTR (%)	NDT	NDTR (%)
B-AS	17	0.628	13	1.343
I-AS	12	0.444	8	0.826
B-C	99	3.660	48	4.959
I-C	109	4.030	58	5.992
B-CH	51	1.885	42	4.339
I-CH	48	1.774	33	3.409
B-DT	130	4.806	67	6.921
I-DT	194	7.172	96	9.917
B-EV	52	1.922	30	3.099
I-EV	12	0.444	10	1.033
B-G	0	0.000	0	0.000
I-G	0	0.000	0	0.000
B-N	33	1.220	7	0.723
I-N	0	0.000	0	0.000
B-OBS	47	1.738	26	2.686
I-OBS	58	2.144	35	3.616
B-R	19	0.702	16	1.653
I-R	14	0.518	13	1.343
B-RA	3	0.111	3	0.310
I-RA	0	0.000	0	0.000
B-T	66	2.440	36	3.719
I-T	36	1.331	28	2.893
B-THER	88	3.253	62	6.405
I-THER	59	2.181	37	3.822
B-V	38	1.405	29	2.996
I-V	62	2.292	18	1.860
O	1,458	53.900	253	26.136
Total	2,705	100	968	100

Table 3: Quantitative analysis of the test dataset

NE	O	OR (%)	DO	DOR (%)
AS	2,488	15.59	1,412	16.14
C	3,887	24.35	2,203	25.18
CH	1,044	6.54	632	7.22
DT	1,519	9.52	883	10.09
EV	793	4.97	331	3.78
G	63	0.39	50	0.57
OBS	217	1.36	166	1.90
N	768	4.81	48	0.55
R	1,766	11.06	1,090	12.46
RA	71	0.45	14	0.16
T	2,041	12.79	1,012	11.57
THER	894	5.60	563	6.44
V	411	2.57	344	3.93
Total	15,962	100.00	8,748	100.00

Table 4: NE Training/Validation Dataset Description

words as out-of-vocabulary, FastText may represent some of them, based on their characters.

For training the FastText model, the following parameters were used: 300 dimensions, skip-gram with negative sampling, minimum count of 5 words, minimum char-gram length of 1, and default settings for the remaining hyperparameters. The skip-gram algorithm (Mikolov et al., 2013a) predicts the surrounding context given the input word, which allows to relate words to their neigh-

NE	O	OR (%)	DO	DOR (%)
AS	17	2.644	14	2.960
C	99	15.397	66	13.953
CH	51	7.932	45	9.514
DT	130	20.218	102	21.564
EV	52	8.087	34	7.188
G	0	0.000	0	0.000
N	33	5.132	7	1.480
OBS	47	7.309	34	7.188
R	19	2.955	17	3.594
RA	3	0.467	3	0.634
T	66	10.264	44	9.302
THER	88	13.686	73	15.433
V	38	5.910	34	7.188
Total	643	100	473	100

Table 5: NE Test Dataset Description

NE	AR (%)	AT	NAT	Total
AS	98.01	1,821	37	1,858
C	94.16	2,323	144	2,467
CH	86.29	428	68	496
DT	93.79	1,193	79	1,272
EV	97.15	375	11	386
G	100.00	27	0	27
N	97.74	259	6	265
OBS	91.11	164	16	180
R	91.68	1,322	120	1,442
RA	91.30	21	2	23
T	96.81	1,273	42	1,315
THER	95.13	605	31	636
V	96.78	331	11	342
O	96.91	8,941	285	9,226
Total	95.73	19,083	852	19,935

Table 6: Agreement Ratios for all NEs and Non-Entity

bors, an important characteristic for NER. The number of dimensions (300) and minimum word count (5) were the same as in the out-of-domain WE model. Minimum char-grams length (1) was used for training the model with all the characters, thus enabling to recognize unknown words. Finally, all the words in the dataset starting with an uppercase character were converted to lowercase, since they represent the same word but in the beginning of a sentence. After preprocessing, only 7,312 tokens occur more than 5 times.

For the out-of-domain WEs, we used a general Portuguese WE model downloaded from the FastText website[2], trained with billions of tokens from Wikipedia and Common Crawl (Grave et al., 2018). As it was trained with a character window of 5 characters, a total of 27 words and 80 lemmas in our dataset do not have an embedding vector in this model. For them, we assign the embedding of the word 'UNK', meaning unknown, but not a Portuguese word, thus not introducing much noise to the embedding datasets. This strategy was followed because simply putting out these words could influence the labelling of the network, as the classification of each word depends on the classification of the others around.

3.3 Model Architecture

Given the current trend on NER and its state-of-the-art results, we adopted a BiLSTM-CRF neural network as our model for this purpose. The architecture used is presented in figure 1. The word embedding step is where all the tokens are converted to their embedding vectors. Lemmas are also converted to their WE vectors and concatenated to the previous vectors. PoS tags, orthographic and morphological features, e.g. first character is uppercase, all characters are uppercase, digit/non-digit were added as well. Afterwards, the embedding vectors are inserted in a BiLSTM layer with one backward layer and one forward layer. The former enables the network to preserve the information from the past to the future, since it analyses the information from the left to the right. The forward layer enables the network to do the inverse of the backward. Together, these types of LSTM improve the prediction of the network, which, this way, understands better the context of each token.

Finally, the output of the BiLSTM layer is inserted in the CRF layer, which enables the network to consider the neighbor tags. In other words, it allows the network to create tag relations, e.g., if a token is tagged with a beginning of NE, the following token is probably the continuation of such NE. This layer is also responsible for not allowing a token to be tagged with an in-NE tag without this NE being started previously.

Adam optimization function (Kingma and Ba, 2014) was used with a learning rate of 0.001. A grid search was not performed here because this study does not focus on the architecture, but on the application of these models to Portuguese.

In order to get the best number of hidden units and dropout percentages for our model, we performed a grid search using 50 training epochs with 10-fold cross validation. As the dataset has a low number of instances, we used a small set of values for the grid search of the number of hidden units [2^3, 2^7]. Keeping the network with a low number of parameters prevents overfitting to the data (Zhang et al., 2016). Furthermore, we used an interval of dropout percentage values from 10%

[2] https://fasttext.cc/docs/en/crawl-vectors.html

to 50%. This hyperparameter allows the network to prevent both overfitting and under-learning (Srivastava et al., 2014). An independent grid search was run for each WE model, because they had been trained in different types of text.

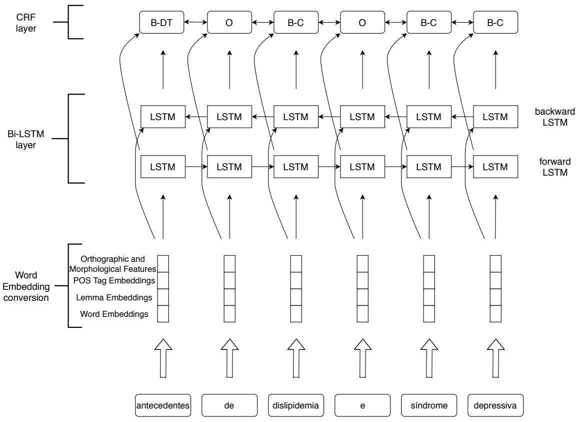

Figure 1: BiLSTM-CRF Neural Network Architecture on the sentence: "antecedentes de dislipidemia e síndrome depressiva" (*history of dyslipidemia and depressive syndrome*)

4 Results and Discussion

According to grid search, the best number of hidden units is 2^6 and 2^5, respectively for the network that uses the in-domain WEs and for the one that uses out-of-domain WEs. The best dropout percentage is 50% for both. This confirms that, for small datasets, the value of each parameter should also be small. Furthermore, the results corroborate that dropout regularization helps avoiding overfitting, since the best results were obtained for high dropout percentage. Validation results for both models and all NE classes are in table 7.

Besides looking at recall and precision, we focus our discussion on the F1-score. Table 7 shows relaxed and strict results. Relaxed or one-point performance measures the performance of the model for each token, while the strict performance considers all occurrences, i.e., one occurrence is well predicted if all its tokens are well predicted too. For example, with the relaxed evaluation, "síndrome depressiva" (*depressive syndrome*) counts as two tokens, i.e, each token's tag is independently compared to its golden tag. With the strict evaluation, if the model fails on a single token's tag, all NE occurrence is considered incorrect.

Results show that the in-domain WE model performs better than the out-of-domain, which is in line with Newman-Griffis and Zirikly (2018). An important reason for this is that the out-of-domain model was not trained with unigrams, leading to the representation of some tokens with the 'UNK' vector, instead of the original token, thus introducing bias. A second reason is that the out-of-domain model was not trained specifically for the clinical domain. Although trained in a much larger collection of text, the out-of-domain model fails to learn clinical relations between different diseases or diagnostic tests, as the in-domain model does. Table 8 shows examples that confirm this fact, e.g. in the in-domain model the word "ECG" is related to three other cardiac diagnostic tests, beyond its extended form, while in the out-of-domain model, it is only related to one more ("ecocardiograma"); or the neighbors of "diabetes" in the in-domain model, which include related diseases (e.g., "dislipidemia" and arterial hypertension ("HTA"), while, in the out-of-domain model, the neighbors of the same word are words that contain it (e.g., "pré-diabetes" and "diabetes.O"). Furthermore, in the out-of-domain model, several words are not related with the clinical domain, as "hemiparasita" (*hemiparasite*) in the "hemiparésia" (*hemiparesis*) example, or words are not related with anything understandable, as in the "poliangeíte" example.

Table 9 has the results for both WE models on the independent test set, and for a CRF model used as a baseline. The CRF was trained in the same dataset, using the same features as the deep learning model, but raw tokens and lemmas, instead of their embeddings. The best hyperparameters of the validation dataset were used for both WEs. This experiment aims to analyze how well the models trained in text from the journal perform on text collected directly from the hospital.

Once again, the in-domain WE model outperformed the out-of-domain model. Average results for this independent dataset are about 10% lower than for the validation dataset. A possible reason for this is that the test set contains some admission notes and patient discharge letters, structured on items (e.g., origin, admission motive) and their description, which is different from the clinical cases in the validation dataset, described in a full paragraph that covers all related information. Furthermore, since they were not published, these texts were written less carefully, and therefore have some orthographic errors.

		Recall		Precision		F1-Score	
WE	**NE**	**Relaxed**	**Strict**	**Relaxed**	**Strict**	**Relaxed**	**Strict**
In-Domain	**mic Avg**	82.34±1.97	74.48±2.37	82.77±1.72	75.25±2.36	82.54±1.61	74.86±2.17
Out-of-Domain		81.63±2.07	73.35±1.57	82.31±1.48	75.06±1.62	81.96±1.50	74.19±1.44
In-Domain	**mac Avg**	79.04±1.99	73.08±3.00	81.06±2.12	75.59±2.77	79.54±1.89	73.87±2.66
Out-of-Domain		77.75±2.84	70.87±3.07	79.71±2.87	73.73±3.42	78.02±2.76	71.58±2.91
In-Domain	**Weighted Avg**	82.34±1.97	74.48±2.37	82.84±1.49	75.23±2.39	82.44±1.59	74.73±2.15
Out-of-Domain		81.63±2.07	73.35±1.57	82.35±1.54	74.82±1.65	81.76±1.59	73.90±1.42

Table 7: 10-fold Cross Validation Results with both WEs

WE	**Word**	**Top-5 Nearest Neighbors**
In-Domain	ECG	ECG-Holter; electrocardiograma; ecodoppler; ecocardiograma ecocardiogramas
Out-of-Domain	ECG	eletrocardiograma; Electrocardiograma; electrocardiograma; ecocardiograma; Ecocardiograma
In-Domain	diabetes	mellitus; dislipidemia; dislipidémia; HTA; diabética
Out-of-Domain	diabetes	diabete; pré-diabetes; Diabetes; Pré-diabetes; diabetes.O
In-Domain	paramnésia	amnésia; amnésico; mnésico; mnésica; desorientação
Out-of-Domain	paramnésia	paramécia; param3; paranóia.; alucinatória; articulatória
In-Domain	polineuropatia	neuropatia; mononeuropatia; axonal; sensitivo-motora; miopatia
Out-of-Domain	polineuropatia	Polineuropatia; polineuropatias; mononeuropatia; polineurite; neuropatia
In-Domain	poliangeíte	ganglonopatia; citopatia; mielopatia; linfoproliferativa; granulomatosa
Out-of-Domain	poliangeíte	CH12CH14CH15CH18CH26CH30CH4DH5DH6DH8DH9DH10DH12DH15DH20DH30DH; estômagoCarbosymagDulcolaxGavisconImodiumIpraaloxLansoylLubentylMaaloxMicrolaxRennieSmectaSpasfon; XIII787980818283848586878889909192Colóquio; AnguloSimulacrosVeículosABCIABSCABTDABTMBRTPBRTSBSRPBSRSLTRGVAMEVAPAVCOCVCOTVEVECIVETAVFCIVGEOVLCIVOPEVPVPMEVPMTVRCIVSAEVSAMVSATVTGCVTPGVTPTVTTFVTTRVTTUVUCIA1; biológicoCaméfitoLigações
In-Domain	hemiparésia	hemiparesia; hemiplegia; hemianopsia; hemianópsia; biparésia
Out-of-Domain	hemiparésia	hemiparéticos; hemiparesia; hemiparasita; hemiplegia; hemiparasitas
In-Domain	artralgias	poliartralgias; algias; mialgias; cervicalgias; lombalgias
Out-of-Domain	artralgias	Artralgias; artralgia; mialgias; Mialgias; Nevralgias

Table 8: Top-5 Nearest Neighbors for both WE models

Average results for the CRF are lower than the average results for both BiLSTM-CRF models. This difference is in line with the results obtained by Chokwijitkul et al. (2018) and Wu et al. (2018). In general, the results of table 9 follow the agreement ratios presented in table 6. Additional Observations and Characterization present the lowest results because they carry too general information easily labelled by the model as a more specific NE (e.g. Condition or Evolution) as explained in section 3.1. Results show low results as well, due to their similarity with Condition, also shown in the examples of section 3.1. Value, Negation, DateTime, Evolution and Anatomical Site show the highest results because they are very specific. Value is related to numbers of therapeutic doses or to the results of diagnostic texts, Negation and Evolution are NEs with many repeated tokens (see tables 2 and 3) and they are highly related to Condition and Results, a characteristic caught by the CRF layer. DateTime is related with time, usually written using the same words and not depending on the author of the text (e.g. training texts contain "aos 60 anos" (*at 60 years old*) and "durante 21 dias" (*during 21 days*) and test texts have "aos 14 anos" (*at 14 years old*) and "durante o período da manhã" (*during the morning*)). Although Anatomical Site has few tokens on the test texts, they are frequent on the training data, which is why results for this NE are high. We were expecting better results for Condition, Test and Therapeutics because they are too specific. This did not happen, and a possible explanation is the different style of writing in the training and testing texts.

Finally, it is important to recall that the Genetics NE is not in the test set, and that the same set has only one token for Negation and Route of Administration, which explains the same relaxed and strict results for these NEs.

5 Conclusion

With this study, we achieved our the three main goals: we gathered and annotated a new dataset for Portuguese clinical text; we applied a BiLSTM-CRF neural network for NER on the previous dataset; we learned a WE model of Portuguese clinical text and compared the performance of the previous approach when using this model and when using general language WEs. The datasets and the learned WE model are publicly available

			Recall		Precision		F1-Score	
Algorithm	**WE**	**NE**	**Relaxed**	**Strict**	**Relaxed**	**Strict**	**Relaxed**	**Strict**
BiLSTM-CRF	**In-Domain**	**AS**	100.00	88.24	80.56	68.18	89.23	76.92
	Out-of-Domain		93.10	88.24	75.00	65.22	83.08	75.00
CRF	-		86.21	70.59	42.37	40.00	56.82	51.06
BiLSTM-CRF	**In-Domain**	**C**	70.19	70.71	59.11	54.26	64.18	61.40
	Out-of-Domain		72.12	68.69	67.87	59.13	69.93	63.55
CRF	-		72.12	61.62	52.63	42.07	60.85	50.00
BiLSTM-CRF	**In-Domain**	**CH**	24.24	23.53	42.11	38.71	30.77	29.27
	Out-of-Domain		21.21	21.57	47.73	45.83	29.37	29.33
CRF	-		15.15	21.57	50.00	44.00	23.26	28.95
BiLSTM-CRF	**In-Domain**	**DT**	85.80	66.15	84.50	71.07	85.15	68.53
	Out-of-Domain		87.64	61.54	82.08	68.38	84.78	64.78
CRF	-		82.41	48.46	76.95	64.29	79.58	55.26
BiLSTM-CRF	**In-Domain**	**EV**	81.25	75.00	82.54	81.25	81.89	78.00
	Out-of-Domain		64.06	53.85	78.85	80.00	70.69	64.37
CRF	-		60.94	51.92	92.86	90.00	73.58	65.85
BiLSTM-CRF	**In-Domain**	**N**	96.97	96.97	88.89	88.89	92.75	92.75
	Out-of-Domain		96.97	96.97	91.43	91.43	94.12	94.12
CRF	-		93.94	93.94	91.18	91.18	92.54	92.54
BiLSTM-CRF	**In-Domain**	**OBS**	17.14	12.77	64.29	40.00	27.07	19.35
	Out-of-Domain		0.95	0.00	33.33	0.00	1.85	0.00
CRF	-		4.76	6.38	100.00	75.00	9.09	11.76
BiLSTM-CRF	**In-Domain**	**R**	63.64	68.42	38.18	44.83	47.73	54.17
	Out-of-Domain		57.58	47.37	45.24	37.50	50.67	41.86
CRF	-		54.55	42.11	19.78	22.22	29.03	29.09
BiLSTM-CRF	**In-Domain**	**RA**	33.33	33.33	50.00	50.00	40.00	40.00
	Out-of-Domain		33.33	33.33	50.00	50.00	40.00	40.00
CRF	-		33.33	33.33	100.00	100.00	50.00	50.00
BiLSTM-CRF	**In-Domain**	**T**	62.75	54.55	68.82	59.02	65.64	56.69
	Out-of-Domain		60.78	48.48	57.41	44.44	59.05	46.38
CRF	-		50.98	34.85	43.70	33.33	47.06	34.07
BiLSTM-CRF	**In-Domain**	**THER**	84.35	67.05	58.49	57.84	69.08	62.11
	Out-of-Domain		79.59	64.77	68.42	62.64	73.58	63.69
CRF	-		69.39	61.36	82.93	80.60	75.56	69.68
BiLSTM-CRF	**In-Domain**	**V**	96.00	84.21	88.07	80.00	91.87	82.05
	Out-of-Domain		89.00	73.68	83.18	66.67	85.99	70.00
CRF	-		86.00	63.16	82.69	63.16	84.31	63.16
BiLSTM-CRF	**In-Domain**	**mic Avg**	70.97	62.36	69.85	63.05	70.41	62.71
	Out-of-Domain		67.68	56.14	72.32	62.03	69.93	58.94
CRF	-		63.43	49.46	63.79	55.11	63.61	52.13
BiLSTM-CRF	**In-Domain**	**mac Avg**	67.97	61.74	67.13	61.17	65.45	60.10
	Out-of-Domain		63.03	54.87	65.04	55.94	61.93	54.42
CRF	-		59.15	49.11	69.59	62.15	56.81	50.12
BiLSTM-CRF	**In-Domain**	**Weighted Avg**	70.97	62.36	69.75	61.91	68.52	61.10
	Out-of-Domain		67.68	56.14	68.20	57.87	66.07	56.26
CRF	-		63.43	49.46	70.07	60.77	61.39	51.31

Table 9: Results of BiLSTM-CRF model using both WEs and of baseline CRF model on independent test set

in our GitHub repository[3]. We hope that making all these resources available for everyone has a positive impact on IE from text written in Portuguese, namely on clinical text.

In-domain WEs were trained with much less text, but lead to higher performance in NER. Although in a different language, this is in line with Newman-Griffis and Zirikly (2018), and confirms that, in the clinical domain, it should be better to train WE models exclusively with clinical texts, even if there is substantially more in-domain text.

The performance of the model in the independent test confirms that it is possible to train models for extracting information from hospital clinical texts without having direct access to them. In other words, IE models trained with public clinical cases extracted from journals are able to extract information from texts never seen before by the model. This is important, given the difficulty to access clinical texts from hospitals.

In order to improve the current results, we plan to make a better parameter optimization and to explore other deep learning architectures, such as those using residual learning (Tran et al., 2017). Furthermore, we aim to increase the datasets used and tackle relation extraction between NEs (Sahu et al., 2016), which would make it easier to summarize clinical reports.

6 Acknowledgements

We would like to thank the Neurology service of the Coimbra University Hospital Centre (CHUC) for providing us access to the testing dataset; and to those responsible for the revision of 30% of the training/validation dataset, namely Paulo Francisco Valente, João Marques, Adriana Leal, Mauro Pinto, Fellipe Allevato and Ricardo Rodrigues. We also acknowledge the financial support of Fundação para a Ciência e a Tecnologia through CISUC (UID/CEC/00326/2019).

[3] https://github.com/fabioacl/PortugueseClinicalNER

References

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Pedro Vitor Quinta de Castro, Nádia Félix Felipe da Silva, and Anderson da Silva Soares. 2018. Portuguese named entity recognition using LSTM-CRF. In *Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2018, Canela, Brazil, September 24-26, 2018, Proceedings*, pages 83–92.

Thanat Chokwijitkul, Anthony Nguyen, Hamed Hassanzadeh, and Siegfried Perez. 2018. Identifying risk factors for heart disease in electronic medical records: A deep learning approach. In *Proceedings of the BioNLP 2018 workshop*, pages 18–27, Melbourne, Australia. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. SemEval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

Liliana Ferreira, António J. S. Teixeira, and João Paulo Cunha. 2010. Information Extraction from Portuguese Hospital Discharge Letters. *VI Jornadas en Technologia del Habla and II Iberian SL Tech Workshop*, (January):39–42.

Liliana da Silva Ferreira. 2011. *Medical Information Extraction in European Portuguese*. Ph.D. thesis, Universidade de Aveiro.

Sherman Folland, Allen C Goodman, and Miron Stano. 2017. Introduction. In *The Economics of Health and Health Care*, 8th edition, chapter 1, pages 29–54. Pearson Prentice Hall Upper Saddle River, NJ.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nathan S. Hartmann, Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings the 11th Brazilian Symposium in Information and Human Language Technology*, STIL 2017.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, Joao Palotti, and Guido Zuccon. 2016. Overview of the CLEF eHealth evaluation lab 2016. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 255–266. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, page 13.

Juliane Klatt, Hinnerk Feldwisch-Drentrup, Matthias Ihle, Vincent Navarro, Markus Neufang, Cesar Teixeira, Claude Adam, Mario Valderrama, Catalina Alvarado-Rojas, Adrien Witon, et al. 2012. The epilepsiae database: An extensive electroencephalography database of epilepsy patients. *Epilepsia*, 53(9):1669–1676.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Thoai Man Luu, Robert Phan, and Rachel Davey. 2018. Clinical Name Entity Recognition Based on Recurrent Neural Networks. *2018 18th International Conference on Computational Science and Applications (ICCSA)*, pages 1–9.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42(5):923–936.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Denis Newman-Griffis and Ayah Zirikly. 2018. Embedding transfer for low-resource medical named entity recognition: A case study on patient mobility. In *Proceedings of the BioNLP 2018 workshop*, pages 1–11, Melbourne, Australia. Association for Computational Linguistics.

John P Pestian, Christopher Brew, Dj J Matykiewicz Pawełand Hovermale, Neil Johnson, K Bretonnel Cohen, and Włodzisław Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics.

Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Ricardo Rodrigues, Hugo Gonçalo Oliveira, and Paulo Gomes. 2018. NLPPort: A Pipeline for Portuguese NLP (Short Paper). In *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*, volume 62 of *OpenAccess Series in Informatics (OASIcs)*, pages 18:1—–18:9, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Sunil Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. 2016. Relation extraction from clinical texts using domain invariant convolutional neural network. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 206–215, Berlin, Germany. Association for Computational Linguistics.

Cicero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China. Association for Computational Linguistics.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826.

Sinapse. 2017a. *Publicações da Sociedade Portuguesa de Neurologia*, volume 17:1. Sociedade Portuguesa de Neurologia, Lisbon.

Sinapse. 2017b. *Publicações da Sociedade Portuguesa de Neurologia*, volume 17:2. Sociedade Portuguesa de Neurologia, Lisbon.

Maria Skeppstedt, Maria Kvist, Gunnar H. Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, 49:148–158.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Amber Stubbs and Özlem Uzuner. 2015. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of biomedical informatics*, 58:S78–S91.

Quan Tran, Andrew MacKinlay, and Antonio Jimeno Yepes. 2017. Named Entity Recognition with Stack Residual LSTM and Trainable Bias Decoding. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 566–575, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Inigo Jauregi Unanue, Ehsan Zare Borzeshi, and Massimo Piccardi. 2017. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *Journal of Biomedical Informatics*, 76(June):102–109.

Özlem Uzuner, Scott L DuVall, Brett R South, and Shuying Shen. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. 2018. Clinical Named Entity Recognition Using Deep Learning Models. *AMIA Annual Symposium Proceedings*, pages 1812–1819.

Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. 2017. A Bidirectional LSTM and Conditional Random Fields Approach to Medical Named Entity Recognition. *International Conference on Advanced Intelligent Systems and Informatics*, pages 355–365.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.