Biomedical Named Entity Recognition in Eight Languages with Zero Code Changes

Veysel Kocaman[1], Gursev Pirge[1] Bunyamin Polat[1] and David Talby [1]
** These authors contributed equally.*
[1] *John Snow Labs Inc. 16192 Coastal Highway, Lewes, DE 19958, USA*

Abstract

Named entity recognition (NER) is one of the most important building blocks of NLP tasks in the medical domain by extracting meaningful chunks from clinical notes and reports, which are then fed to downstream tasks like assertion status detection, entity resolution, relation extraction, and de-identification. Due to the growing volume of healthcare data in unstructured format, an increasingly important challenge is providing high accuracy implementations of state-of-the-art deep learning (DL) algorithms at scale. On the other hand, when it comes to low-resource languages, collecting high quality annotated data sets in the biomedical domain is still a big challenge. In this study, we train production-grade biomedical NER models on eight different biomedical datasets published within the LivingNER competition [1]. Transformer based Bert for token classification and BiLSTM-CNN-Char based NER algorithms from Spark NLP library are utilized during this study and we trained 28 different NER models in total with decent accuracies (0.9243 F1 test score in Spanish) without changing the underlying DL architecture. The trained models are available within a production-grade code base as part of the Spark NLP library; can scale up for training and inference in any Spark cluster; has GPU support and libraries for popular programming languages such as Python, R, Scala and Java.

Keywords

Named entity recognition, NER, deep learning, NLP, Spark NLP

1. Introduction

Natural language processing (NLP) can be defined as an important part of artificial intelligence and focuses on the automatic analysis, representation and understanding of the human language [2]. Considering that manual abstraction is a highly expensive, time consuming and error prone process, there has been a growing trend in NLP applications in the clinical and biomedical domain to automate the abstraction process. Common use cases of NLP include question answering, paraphrasing, summarizing, sentiment analysis, natural language BI, language modelling, and disambiguation [3].

NLP is always just a part of a bigger data processing pipeline and due to the nontrivial steps involved in this process, there is a growing need for an all-in-one solution to ease the burden of text preprocessing at large scale and connecting the dots between various steps of solving a data science problem. A good NLP library should be able to correctly transform the free text into structured features and let the users train their own NLP models that are easily fed into the downstream machine learning (ML) or deep learning (DL) pipelines with no hassle [3].

A primary building block in such text mining systems is named entity recognition (NER) - which is regarded as a critical precursor for question answering, topic modelling, information retrieval, etc. [4]. In the medical domain, NER recognizes the first meaningful chunks out of a clinical note, which are

IberLEF 2022, September 2022, A Coruña, Spain.
EMAIL: veysel@johnsnowlabs.com (A. 1); gursev@johnsnowlabs.com (A. 2); bunyamin@johnsnowlabs.com (A. 3); david@johnsnowlabs.com (A. 4)
ORCID: 0000-0002-0065-6478 (A. 1); 0000-0002-0641-5392 (A. 2); 0000-0002-1226-9969 (A. 3); 0000-0003-2782-5478 (A. 4)

then fed down the processing pipeline as an input to subsequent downstream tasks such as clinical assertion status detection [5], clinical entity resolution [6] and de-identification of sensitive data [7]. However, segmentation of clinical and drug entities is a difficult task in biomedical NER systems because of complex orthographic structures of named entities [8].

In this study, we share our results on the NER track of LivingNER 1, a contest organized/sponsored by The Spanish National Bioinformatics Institute (INB) that focuses specifically on the automatic detection of species mentions (humans, plants, animals, insects, pathogens), as well as their normalization to species taxonomy concepts. The original contest has 3 tracks: (i) LivingNER-Species NER, (ii) LivingNER-Species Norm and (iii) LivingNER-Clinical IMPACT.

As a John Snow Labs team, we only joined the first task (LivingNER-Species NER) and trained several NER models on Spanish clinical case reports. Our best model achieved 0.9243 F1 score on the test set (0.9645 on NER only HUMAN and 0.8964 on NER only SPECIES). Within the same contest, the organizers also shared the translated version of the same dataset into seven languages (English, Catalan, French, Portuguese, Galatia, Italian and Romanian) with no test set. Nevertheless, we trained NER models on these datasets and achieved decent (validation) F1 scores that are higher than 0.8. We used various embeddings models for each NER model and observed that using BERT language model in the embeddings stage of NER architecture (BiLSTM-CNN-Char) produced better results compared to Bert for Token Classification method (end to end NER architecture used in transformers) on validation set while Bert for Token Classification method performed better on test sets.

2. Background

2.1. Spark NLP library

All the experiments are run within Spark NLP library, a popular open-source NLP library that has been downloaded more than 25 million times so far. Spark NLP library is developed to be a single unified solution for all the NLP tasks and is the only library that can scale up for training and inference in any Spark cluster, take advantage of transfer learning and implementing the latest and greatest algorithms and models in NLP research, and deliver a mission-critical, enterprise- grade solutions at the same time. It is an open-source natural language processing library, built on top of Apache Spark and Spark ML. It provides an easy API to integrate with ML pipelines and it is commercially supported by John Snow Labs Inc, an award-winning healthcare AI and NLP company based in the USA [3].

Spark NLP library has two versions: Open source and enterprise. The open-source version has all the features and components that could be expected from any NLP library, using the latest DL frameworks and research trends. Enterprise library is licensed (free for academic purposes) and designed towards solving real world problems in the healthcare domain and extends the open-source version. The licensed version has the following modules to help researchers and data practitioners in various means: Named entity recognition (NER), assertion status (negativity scope) detection, relation extraction, entity resolution (SNOMED, RxNorm, ICD10 etc.), clinical spell checking, contextual parser, text2SQL, deidentification and obfuscation [3].

In a previous study [9], it was shown through extensive experiments that NER module in Spark NLP library exceeds the biomedical NER benchmarks reported by Stanza [13] in 7 out of 8 benchmark datasets and in every dataset reported by SciSpacy [14] without using heavy contextual embeddings like BERT. Using the modified version of the well-known BiLSTM-CNN-Char NER architecture [10] into Spark environment, even with a general-purpose GloVe embeddings (GloVe6B) and with no lexical features, state- of-the-art results in biomedical domain were achieved and better results than Stanza in 4 out of 8 benchmark datasets were produced.

2.2. LivingNER contest

Organisms/species have been scarcely featured in NLP studies to date, particularly for non-English content. LivingNER is a contest organized/sponsored by The Spanish National Bioinformatics Institute (INB) that focuses specifically on the automatic detection of species mentions (humans, plants, animals, insects, pathogens), as well as their normalization to species taxonomy concepts. LivingNER-Species NER track (Species mention entity recognition) is part of the contest. Given a collection of plain text clinical case report documents, participants are expected to return the exact character offsets of all species mentioned, both human and non-human. The annotation of species or living organisms is critical to scientific disciplines like medicine, biology, ecology/biodiversity, nutrition and agriculture. The distribution of the entities for each dataset can be found in Table 1.

Table 1
The distribution of the entities for each dataset in LivingNER-Species NER track.

Label	Training	Validation	Test
Human	7007	3289	79837
Species	9090	3817	40291

This study involves detection of species mentions by using the Spark NLP library and four different word embeddings in Spanish, with the aim of identifying the species in biomedical literature generated in Spanish. As a comparison, transformer-based Bert For Token Classification (B4TC) NER method was also used. Results were investigated to understand the effects of using different embeddings on the performance of the models, as each embedding is a distributed vector representation technique to capture information of a word in the text.

2.3. Named Entity Recognition in Spark NLP

The deep neural network architecture for named entity recognition in Spark NLP is based on the BiLSTM-CNN-Char framework [11] and it is a modified version of the architecture proposed by Chiu et.al. [10]. It may be defined as a neural network architecture that automatically detects word and character-level features using a hybrid bidirectional LSTM and CNN architecture, eliminating the need for most feature engineering steps. The detailed architecture of the framework in the original paper is illustrated in Figure 1.

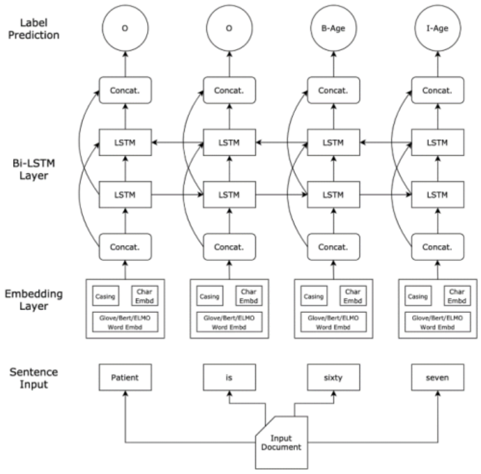
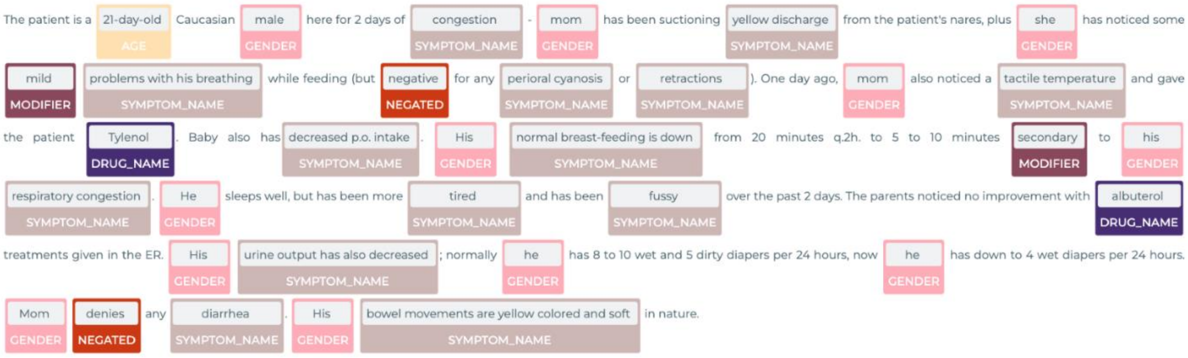


Figure 1: Overview of the original BiLSTM-CNN-Char architecture proposed by Chiu et al. [10].

Sample predictions from one of the pre-trained clinical NER models from Spark NLP for Healthcare library can be seen in Figures 2 and 3.



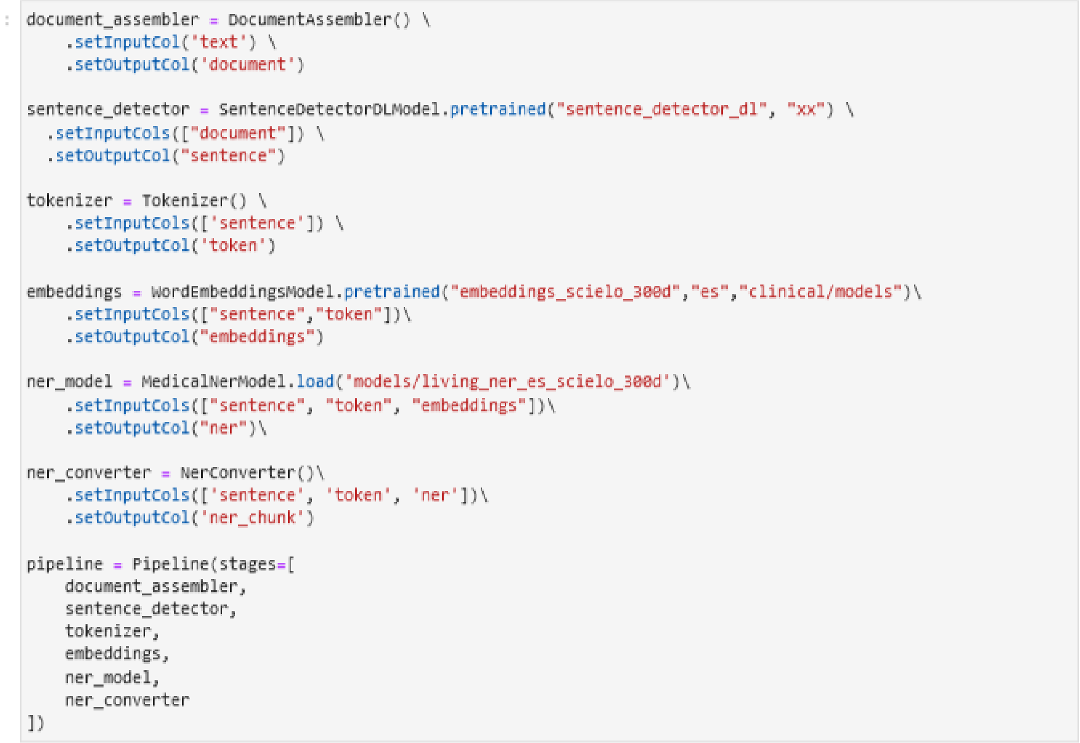
The patient is a 21-day-old Caucasian male here for 2 days of congestion - mom has been suctioning yellow discharge from the patient's nares, plus she has noticed some mild problems with his breathing while feeding (but negative for any perioral cyanosis or retractions). One day ago, mom also noticed a tactile temperature and gave the patient Tylenol. Baby also has decreased p.o. intake His normal breast-feeding is down from 20 minutes q.2h. to 5 to 10 minutes secondary to his respiratory congestion He sleeps well, but has been more tired and has been fussy over the past 2 days. The parents noticed no improvement with albuterol treatments given in the ER. His urine output has also decreased; normally he has 8 to 10 wet and 5 dirty diapers per 24 hours, now he has down to 4 wet diapers per 24 hours. Mom denies any diarrhea His bowel movements are yellow colored and soft in nature.

Figure 2: Sample clinical entities predicted by a clinical NER model trained on various datasets. There are more than 100 pretrained NER models in Spark NLP Enterprise edition [3].

In Spark NLP, this architecture is implemented using TensorFlow, and has been heavily optimized for accuracy, speed, scalability, and memory utilization. This setup has been tightly integrated with Apache Spark to let the driver node run the entire training using all the available cores on the driver node. There is a CuDA version of each TensorFlow component to enable training models on GPU when available. The Spark NLP provides open-source APIs in Python, Java, Scala, and R - so that users do not need to be aware of the underlying implementation details (TensorFlow, Spark, etc.) in order to use it.

Figure 3: Some clinical entities covered by the pretrained NER models in Spark NLP Enterprise edition [3]. There are more than 400 clinical entities covered by more than 100 NER models.

NER systems are usually a part of an end-to-end NLP pipeline (Figure 4) through which the text is fed and then several text preprocessing steps are applied. Since the DL algorithm we implement is sentence-wise, and the features (embeddings and casing) are token-wise, sentence splitting, and tokenization are the most important steps leading to a better accuracy. Using a DL based sentence detector module [12] and a highly customizable rule-based tokenizer in Spark NLP, we ensured that the generated features are more informative. If a token, labelled as an entity in training set cannot be isolated from its appended chars and punctuations, it might be treated as an out of vocabulary word while getting the embeddings, hence zero embeddings, and this harms the learning process that works best when all of the features exist for each token (embeddings, casing and char features). The Spark NLP pipeline for a NER prediction process (MedicalNerModel) can be seen in Figure 4.



```
document_assembler = DocumentAssembler() \
    .setInputCol('text') \
    .setOutputCol('document')

sentence_detector = SentenceDetectorDLModel.pretrained("sentence_detector_dl", "xx") \
  .setInputCols(["document"]) \
  .setOutputCol("sentence")

tokenizer = Tokenizer() \
    .setInputCols(['sentence']) \
    .setOutputCol('token')

embeddings = WordEmbeddingsModel.pretrained("embeddings_scielo_300d","es","clinical/models")\
    .setInputCols(["sentence","token"])\
    .setOutputCol("embeddings")

ner_model = MedicalNerModel.load('models/living_ner_es_scielo_300d')\
    .setInputCols(["sentence", "token", "embeddings"])\
    .setOutputCol("ner")\

ner_converter = NerConverter()\
    .setInputCols(['sentence', 'token', 'ner'])\
    .setOutputCol('ner_chunk')

pipeline = Pipeline(stages=[
    document_assembler,
    sentence_detector,
    tokenizer,
    embeddings,
    ner_model,
    ner_converter
])
```

Figure 4: Spark NLP prediction pipeline, using a pre-trained model in the MedicalNerModel annotator and 300-dimensional embeddings (Scielo 300d) Spanish embeddings.

3. Implementation Details

Biomedical NER datasets were provided by LivingNER [15], which is a contest organized/ sponsored by The Spanish National Bioinformatics Institute (INB) [16]. The dataset involved “HUMAN”, “SPECIES” or “O” entities, with B- tags the beginning of an entity and I- tags the words that are inside the entity. All other words not describing entities of interest are tagged as ‘O’.

As a starting point, models were trained by using Spark NLP MedicalNerApproach [17], which is an NER annotator, that allows to train generic NER models based on BiLSTM-CNN-Char architecture with different embeddings. The embeddings used for training in Spanish language are shown below:

- Scielo Embeddings in Spanish 300d (embeddings_scielo_300d) [18]
- Word2Vec Embeddings in Spanish 300d (w2v_cc_300d) [19]
- Roberta Clinical Word Embeddings (Spanish) (roberta_base_biomedical) [20]
- Spanish BERT Base Cased Embeddings (bert_base_cased) [21]

For other languages, following embeddings were used during training (Table 2):

Table 2
Embeddings used for training the documents in other languages.

Language	Embeddings
English	clinical_emb
	biobert_emb
Galatian	w2v_300d
French	w2v_300d
	bert_base
Portuguese	w2v_300d
	roberta_base
	biobert_emb_all
	biobert_emb_biomed
	withbert_bert_base
Italian	bert_base_italian_xxl_cased
	bert_base_it_cased
	w2v_300d
Romanian	w2v_300d
	bert_base
Catalan	w2v_300d

To improve the accuracy of the models, the following parameters [22] of the *MedicalNerApproach* were used. Range of parameters that were tested and the optimum values are presented in Table 3. In our models, the optimum number of epochs and Batch Size were 32. Optimum values for the Learning Rate were 0.003 and 0.0003, depending on the embeddings used for training.

Once the pre-trained models were loaded within a pipeline, those models were used for prediction. Another Spark NLP NER annotator, MedicalNerModel, was used to measure the performances of the models on the test dataset. To measure the efficiency of this method, we compared the performance of the BertForTokenClassification NER model with the models we trained on Spark NLP.

Table 3
Detailed information about the parameters used for improving the accuracy during training of models.

Model/Embedding	Parameter	Tested Range	Optimum Value
w2v_cc_300d	Epochs	8 - 64	32
Scielo	Batch Size	8 - 64	32
	Learning Rate	0.01 - 0.0003	0.003
	Dropout Rate	0.3 – 0.7	0.3
bert_base	Epochs	8 - 64	32
roberta_base	Batch Size	8 - 64	32
	Learning Rate	0.0001 - 0.0005	0.0003
	Dropout Rate	0.3 – 0.7	0.3
B4TC	Epochs	10 – 32	20
	Batch Size	8 – 128	64
	Learning Rate	0.0001 - 0.00005	0.00003
	Dropout Rate		

4. Experimental Results

Using the official train sets from the contest, we trained NER models and obtained metrics on the same test sets used in the challenges. The results for Spanish datasets can be seen in Table 4, where all the tested configurations obtained decent accuracies over the validation dataset. On the test dataset, the models' performances varied slightly with MedicalNerModel with Spanish BERT Base Cased Embeddings while BertForTokenClassification's performance was slightly better than our MedicalNER models.

Considering the F1 scores for the models trained by Spark NLP's MedicalNerApproach, all of our models produced better performances (ranging between 0.9714 and 0.9389) during training on the Spanish validation dataset compared to BertForTokenClassification model (0.9066). Also, Precision and Recall values in our models were quite close (average difference around 0.01), but there was a difference of 0.05 BertForTokenClassification Precision and Recall values.

The significant result worth mentioning is, even though the MedicalNerModel with word2vec embeddings in Spanish 300d produced the best scores in training, there was a significant drop in the model's performance on the test dataset (F1 scores of 0.9714 for training and 0.7037 for test dataset). The reason may be due to low coverage of our word vectors (large number of out of vocabulary tokens) in the test set. Another reason would be having two times higher SPECIES entities in the test set compared to HUMAN entities (Table 1). Having F1 scores for the HUMAN label mostly around 0.96, whereas it is 0.86 for the SPECIES label can also be explained by the same skewness in the test set. We will be able to analyze this better once the test set is officially shared.

Table 4

Comparison of pre-trained models' performances with various embeddings on training and testing datasets in Spanish. All scores are Precision, Recall and micro-averaged test F1 excluding O's.

Model	Validation Metrics			Test NER			Test – Species NER			Test - Human NER		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
B4TC	0.8805	0.9344	0.9066	0.916	0.9327	0.9243	0.8882	0.9047	0.8964	0.9586	0.9705	0.9645
MedicalNerModel (bert_base)	0.9711	0.9633	0.9671	0.8965	0.9212	0.9087	0.8552	0.8861	0.8704	0.9646	0.9686	0.9666
MedicalNerModel (roberta_base)	0.9724	0.9692	0.9708	0.8803	0.9091	0.8944	0.8367	0.8739	0.8549	0.9514	0.9565	0.9539
MedicalNerModel (scielo_300d)	0.9374	0.9405	0.9389	0.8965	0.9137	0.905	0.8563	0.8749	0.8655	0.9578	0.966	0.9619
MedicalNerModel (w2v_300d)	0.9707	0.9721	0.9714	0.6401	0.7814	0.7037	0.5364	0.7137	0.6125	0.8445	0.8727	0.8584

The dataset was originally written in Spanish and then translated with a neural machine translation system by the content organizers to seven other languages. This translation may be the cause for the metrics for other languages (Table 5) being lower than the original Spanish dataset (Table 4).

In all the other languages, at least one version of our MedicalNER models performed better than BertForTokenClassification versions. Table 5 shows that the highest F1 score for our pre-trained models was obtained by using BioBert embeddings for the English dataset (0.8589), closely followed by the model that used Italian Bert Base Cased Embeddings (0.8500). Highest score by BertForTokenClassification was 0.8110 for the English dataset and lower scores for other datasets. BertForTokenClassification has no embeddings available in Catalan language, whereas our model with Word2Vec Embeddings in Catalan (300d) produced an F1 Score of 0.8258.

Table 5
Comparison of pre-trained models' performances with various embeddings on the training dataset in various languages. All scores are Precision, Recall and micro-averaged test F1 excluding O's.

Language	Model and Embeddings	Precision	Recall	F1
English	B4TC	0.73	0.9121	0.811
	MedicalNER with clinical_emb	0.71	0.92	0.82
	MedicalNER with biobert_emb	0.8566	0.9058	0.8589
Galatian	B4TC	0.5179	0.8582	0.6459
	MedicalNER with w2v_300_d	0.6423	0.9121	0.7538
French	B4TC	0.5248	0.9115	0.6661
	MedicalNER with w2v_300_d	0.7064	0.9235	0.8005
	MedicalNER with bert_base	0.7108	0.8859	0.7887
Portuguese	B4TC	0.6674	0.8976	0.7656
	MedicalNER with w2v_300_d	0.6226	0.9018	0.7366
	MedicalNER with roberta_base	0.6304	0.8575	0.7266
	MedicalNER with biobert_emb_all	0.7765	0.8650	0.8184
	MedicalNER with biobert_emb_biomed	0.7671	0.8487	0.8058
	MedicalNER with bert_base	0.7773	0.8335	0.8044
Italian	B4TC	0.6674	0.8976	0.7656
	MedicalNER with bert_base_italian_xxl_cased	0.80	0.90	0.85
	MedicalNER with bert_base_it_cased	0.73	0.91	0.81
	MedicalNER with w2v_300_d	0.68	0.92	0.79
Romanian	B4TC	0.5248	0.9115	0.6661
	MedicalNER with w2v_300_d	0.1816	0.3383	0.2363
	MedicalNER with bert_base	0.7838	0.8714	0.8253
Catalan	MedicalNER with w2v_300_d	0.7424	0.9303	0.8258

Most of the models produced quite similar Precision and Recall values for the Spanish dataset. On the other hand, for the translated datasets in eight languages, Recall values were always significantly greater than Precision values.

The evolution of accuracy and loss during training is given in Figure 5. The training rapidly improves over the first 10 epochs (accuracy becomes greater than 90%). Afterwards, the improvements are small for the rest of the epochs. During the model training with Scielo Embeddings in Spanish 300d , the improvement was much faster (in the first 5 epochs), but the error was not stable until the end of training.

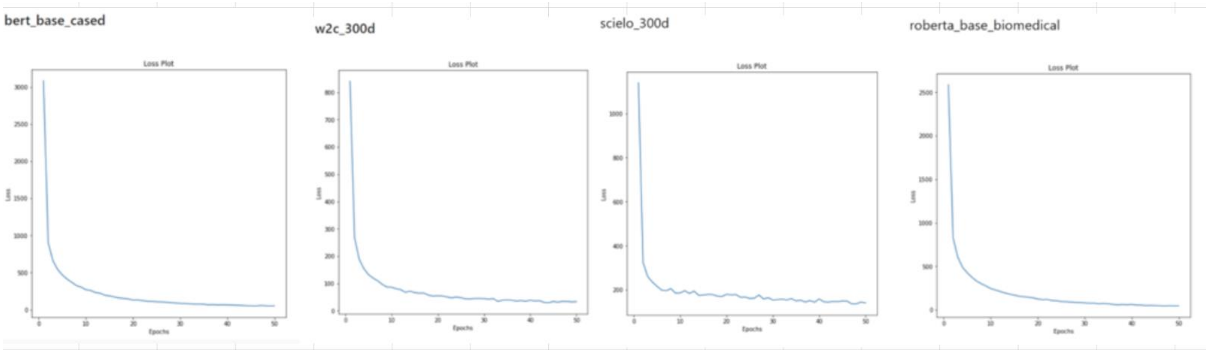


Figure 5: Progress of loss function over 50 training epochs.

5. Conclusion

Despite the growing interest and groundbreaking advances in NLP research and NER systems, easy to use production ready models and tools are scarce in the clinical and biomedical domain and it is one of the major obstacles for clinical NLP researchers to implement the latest algorithms into their workflow and start using immediately.

In this study, we show through experiments on LivingNER contest datasets in eight languages that the BiLSTM-CNN-Char based NER module (MedicalNerApproach) of the Spark NLP library requires no handcrafted features or task-specific resources, achieves decent scores on biomedical datasets and exceeds memory-intensive Bert based BertForTokenClassification models. For all the languages, we trained one BertForTokenClassification and 20 MedicalNER models in total; summing up to 28 different NER models that all can be used as a pretrained NER model out of the box within Spark NLP for Healthcare library.

Considering the F1 scores for the models trained by MedicalNerApproach, all the models produced better performances (ranging between 0.9714 and 0.9389) during training on the Spanish validation dataset compared to BertForTokenClassification model (0.9066). Even though the MedicalNerModel with word2vec embeddings in Spanish 300d produced the best scores on validation set, there was a significant drop in the model's performance on the test dataset (F1 scores of 0.9714 for training and 0.7037 for test dataset) probably due to skewness between entity classes in the test set. In all the other languages, at least one version of our MedicalNER models performed better than BertForTokenClassification versions.

Spark NLP's NER module can also be extended to other spoken languages with zero code changes as long as the respective embeddings or language model exists and can scale up in Spark clusters. In addition, this model is available within a production-grade code base as part of the Spark NLP library; can scale up for training and inference in any Spark cluster; has GPU support and libraries for popular programming languages such as Python, R, Scala and Java.

6. References

[1] IberLEF, Natural Language Processing (NLP) systems in Spanish and other Iberian languages, 2022. URL: https://sites.google.com/view/iberlef2022.

[2] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," IEEE Computational Intelligence Magazine, 8, 2022.

[3] V.Kocaman and D. Talby, "Spark NLP: Natural Language Understanding at Scale", arXiv:2101.10848v1 [cs.CL], 26 Jan 2021.

[4] Li, J., Sun, A., Han, J. and Li, C., "A Survey on Deep Learning for Named Entity Recognition", IEEE Transactions on Knowledge and Data Engineering 34(1), January 2022.

[5] Uzuner, ̇O.; South, B. R.; Shen, S.; and DuVall, S. L. 2011, “2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text”, Journal of the American Medical Informatics Association 18(5): 552–556.

[6] Tzitzivacos, D. 2007, “International Classification of Diseases 10th edition (ICD-10):: main article”, CME: Your SA Journal of CPD 25(1): 8–10.

[7] Uzuner, O., Luo, Y. and Szolovits, P. 2007, “Evaluating the state-of-the-art in automatic de-identification”, Journal of the American Medical Informatics Association 14(5): 550–563.

[8] Liu, S., Tang, B., Chen, Q. and Wang, X, 2015, “Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries”, Information 6(4): 848–865.

[9] V. Kocaman and D. Talby, Biomedical named entity recognition at scale, arXiv preprint arXiv:2011.06315.

[10] J. P. Chiu and E. Nichols, Named entity recognition with bidirectional lstm-cnns, Transactions of the Association for Computational Linguistics 4 (2016) 357–370.

[11] V. Kocaman and D. Talby, “Improving Clinical Document Understanding on COVID-19 Research with Spark NLP”, arXiv preprint arXiv:2012.04005v1.

[12] Schweter, S., Ahmed, S.: Deep-eos: General-purpose neural networks for sentence boundary detection. In: KONVENS (2019).

[13] Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D., 2020. “Stanza: A Python natural language processing toolkit for many human languages”, Stanford NLP, DOI:10.18653/v1/2020.acl-demos.14.

[14] Neumann, M., King, D., Beltagy, I. and Ammar, W., 2019. “SciSpaCy: fast and robust models for biomedical natural language processing”. Proceedings of the 18th BioNLP Workshop and Shared Task, 2019 (8).

[15] LivingNER: Named entity recognition, normalization & classification of species, pathogens and food, 2022. URL: https://temu.bsc.es/livingner/.

[16] The Spanish National Bioinformatics Institute, URL: https://inb-elixir.es/about-inb.

[17] John Snow Labs, Medical NER Annotator, URL: https://nlp.johnsnowlabs.com/docs/en/licensed_annotators #medicalner.

[18] John Snow Labs, Embeddings Scielo 300 dims, URL: https://nlp.johnsnowlabs.com/2020/05/26/embeddings_scielo_300d_es.html.

[19] John Snow Labs, Word2Vec Embeddings in Spanish (300d), URL: https://nlp.johnsnowlabs.com/2022/03/16/w2v_cc_300d_es_3_0.html.

[20] John Snow Labs, Roberta Clinical Word Embeddings (Spanish), URL: https://nlp.johnsnowlabs.com/2021/11/01/roberta_base_biomedical_es.html.

[21] John Snow Labs, Spanish BERT Base Cased Embeddings, URL: https://nlp.johnsnowlabs.com/2021/09/07/bert_base_cased_es.html.

[22] John Snow Labs, Medical NER Approach Annotator, URL: https://nlp.johnsnowlabs.com/licensed/api/python/reference/autosummary/sparknlp_jsl.annotator.MedicalNerApproach.html.