IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 23, 2020, accepted August 10, 2020, date of publication August 24, 2020, date of current version September 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018688

Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives

LILIYA AKHTYAMOVA[1], PALOMA MARTÍNEZ[2], KARIN VERSPOOR[3,4], AND JOHN CARDIFF[1]

[1]Department of Computing, Technological University Dublin (Tallaght Campus), D06 F793 Dublin, Ireland
[2]Computer Science Department, Carlos III University of Madrid, Madrid 28903, Spain
[3]School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC 3010, Australia
[4]Medical School, The University of Melbourne, Melbourne, VIC 3010, Australia

Corresponding author: Liliya Akhtyamova (akhtyamova@phystech.edu)

This work was supported in part by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, DeepEMR project, under Grant, TIN2017-87548-C2-1-R. The work of Karin Verspoor was supported by the University of Melbourne, through a Study Leave grant. The work of Liliya Akhtyamova was supported by the Technological University Dublin as part of a President's Research Award, and a traineeship in UC3M Spain, through an Erasmus+ grant.

ABSTRACT In the Big Data era, there is an increasing need to fully exploit and analyze the huge quantity of information available about health. Natural Language Processing (NLP) technologies can contribute by extracting relevant information from unstructured data contained in Electronic Health Records (EHR) such as clinical notes, patients' discharge summaries and radiology reports. The extracted information can help in health-related decision making processes. The Named Entity Recognition (NER) task, which detects important concepts in texts (e.g., diseases, symptoms, drugs, etc.), is crucial in the information extraction process yet has received little attention in languages other than English. In this work, we develop a deep learning-based NLP pipeline for biomedical entity extraction in Spanish clinical narratives. We explore the use of contextualized word embeddings, which incorporate context variation into word representations, to enhance named entity recognition in Spanish language clinical text, particularly of pharmacological substances, compounds, and proteins. Various combinations of word and sense embeddings were tested on the evaluation corpus of the PharmacoNER 2019 task, the Spanish Clinical Case Corpus (SPACCC). This data set consists of clinical case sections extracted from open access Spanish-language medical publications. Our study shows that our deep-learning-based system with domain-specific contextualized embeddings coupled with stacking of complementary embeddings yields superior performance over a system with integrated standard and general-domain word embeddings. With this system, we achieve performance competitive with the state-of-the-art.

INDEX TERMS Clinical case narratives, contextualized word embeddings, deep learning, language representations, named entity recognition, natural language processing, spanish language.

I. BACKGROUND

Currently, most research in Natural Language Processing (NLP) is focused on English language texts, while text written in different languages is often left unexplored; this has been particularly true in the domain of biomedicine. Given the amount of data produced every year by biomedical experts, doctors, and patients in non-English speaking countries, this represents a significant missed opportunity.

The PharmacoNER 2019 challenge [1] aimed to close the gap in named entity recognition (NER) of biomedical concepts in a corpus of Spanish clinical case narratives. The corpus includes annotations of clinical terminology, chemical and protein entities.

Extraction of biomedical entities from these narratives is relevant to a number of NLP tasks such as adverse drug and drug-drug interaction extraction [2], [3], biomedical concept normalization, knowledge base population [4], and question answering [5].

Recent developments in NLP have shown the advantage of Neural Network (NN)-based methods, particularly those

The associate editor coordinating the review of this manuscript and approving it for publication was Shahzad Mumtaz.

based on Deep Learning, over traditional Machine Learning (ML) algorithms. However, beyond the development of new NN-based methods, researchers have started to explore the impact of improved strategies for the representation of text information provided as input to both NN-based and other ML methods.

Starting from Bag of Words (BoW) representations, word pre-processing has evolved to include more sophisticated word representations such as `word2vec` word embeddings [6], `Glove` [7] and `FastText` [8] embeddings, with the latter two able to capture the subword information from texts. Applied in a range of different NLP tasks, methods using word embeddings have led to significant breakthroughs in model performance for biomedical NER tasks where limited training data is available [9].

Further advances to text preprocessing have been proposed based on language models, that give a word a different embedding vector based on its usage context. The embedding function is trained either from a language modeling perspective [10] or based on recovering masked parts of tokens [11]. The downstream tasks which incorporate these embeddings are considered to be learned in a semi-supervised manner because they benefit from large amounts of unlabeled data [12], [13].

Language representation models can be further applied with or without fine-tuning to problems arising in different domains.[1] The approach of learning on one dataset and applying the model to another dataset is called *Transfer Learning*.

Among recently introduced contextualized embeddings are Semi-supervised Sequence Learning [14], `ELMo` [10], `ULMFiT` [13], the OpenAI transformer [15], the Transformer [16], `BERT` [11] and `Flair` [17].

In our experiments, we explore the use of both `Flair` and `BERT` contextualized embeddings as they have been shown to outperform other types of embeddings on a variety of sequence labeling tasks [11], [17].

In addition to pre-trained domain-specific Spanish `FastText` embeddings [18], we generate domain-specific Spanish contextualized embeddings by pre-training language representation models using the corpus retrieved from the Scientific Electronic Library Online (SciELO) website.[2] The clinical case narrative data from the publications there was used to construct the PharmacoNER dataset. To the best of our knowledge, these are the first contextualized embeddings for Spanish clinical texts made available to a wide audience. The large corpus of more than one billion sentences from SciELO we make available is itself a valuable resource.

This paper extends and deepens a preliminary version of our experiments, which are described in [19]. In particular, we add experiments using the Flair framework which outperform our previous results obtained with the Bert model. We also experiment with word embedding stacking approaches, further improving the results we obtained on the PharmacoNER corpus.

The contributions described in this paper are as follows: (1) we retrieve task-specific corpora for training; (2) we construct task-specific contextualized word embeddings from scratch based on `Flair` and `BERT` architectures; (3) we compare model performances based on constructed word embeddings, explore how these may be combined with other types of embeddings, and compare these with the standard embeddings, producing new baselines; and (4) we conduct an extensive error analysis checking the source of errors for different models.

The pretrained weights for `Flair` and `BERT` models, as well as the SciELO corpora used for their training are made publicly available in a Google Drive repository.[3]

A. BIOMEDICAL ENTITY EXTRACTION APPROACHES

Simple approaches to biomedical NER which sometimes give surprisingly good results have made use of rules or dictionaries.

For example, Eftimov *et al.* [20] built a set of regular expressions to extract evidence-based dietary recommendations from scientific publications and websites. They first detected target mentions in textual data and then extracted them using the rule-based technique.

Various strategies for dictionary lookup have also been shown to be effective [21]. Such approaches leverage biomedical terminology resources or ontologies, and are particularly relevant for biomedical NER where named entities often correspond to fine-grained domain-specific concepts.

However, with the development of automatic NLP methods, these methods are rarely applied on their own to solve NER tasks, but rather are used to generate features to feed ML and deep learning (DL) models. For example, in a recent Meddocan challenge on Spanish medical document anonymization [22], rule-based techniques were actively utilized in ML and DL methods to identify patients' email addresses, locations, phone numbers, etc. In addition, participants in the challenge used domain- and language-specific gazetteers and Brown clusters derived through unsupervised ML. For example, Perez *et al.* [23] concluded that Brown clusters and gazetteers played a significant role in ML system performance. Further, Lopez at al. [24] tested both ML and rule-based approaches and concluded that a hybrid of the two gives the best result.

Lee *et al.* [25] solve the problem of biomedical NER in two steps, firstly by discovering entities' boundaries using Support Vector Machines (SVM) techniques and then further applying an ontology-based hierarchical classification method to classify identified entities. Their system got promising results 66.7 F-score on GENIA corpus [26].

Early work on machine learning-based NER includes such techniques as reranking relying on kernels [27] as well as pure feature processing [28]. Kernel-based methods for

[1] https://ai.googleblog.com/2019/07/advancing-semi-supervised-learning-with.html

[2] https://scielo.org/es

[3] http://dx.doi.org/10.17632/vf6jmvz83b.2

entity extraction such as SVM utilized in numerous papers [29]–[31] overall became popular methods for extracting entities from texts including biomedical texts [32]. In the latter paper, the authors examined different kernel functions for the problem of biomedical NER and concluded that tree-based kernel is more capable of entity extraction.

Current state-of-the-art methods for NER are based on NN architectures, in particular, DL convolutional NNs (CNN) and recurrent NNs (RNN). Transfer learning approaches, in particular the use of pre-trained contextualized word embeddings, have augmented performance of these methods, giving strong results in a number of downstream tasks.

For example, in the Meddocan shared task the best result was achieved by a system which utilized pretrained contextualized `Flair` embeddings fed into a simple RNN model. However, while dealing with more complex biomedical NER problems including long, discontinuous, overlapping entities, hybrid approaches show the best results. Li *et al.* [33] integrated KB embeddings in their tree-structured LSTM framework, achieving approximately 3 point gain in F-score.

Related to this, contextualized word embeddings together with part-of-speech (PoS) tags were examined for Bulgarian NER [34] showing sizeable improvements over the state-of-the-art. In another work, a combination of different types of contextualized embeddings was explored over English biomedical literature corpora [35]. The best results were obtained when combining `ELMo` and `Flair` word embeddings. Another relevant work includes the extraction of adverse drug events on 2018 N2C2 shared task corpus [36]. The authors experimented with the off-the-shelf `Flair` NER framework and kernel-based methods and concluded that a neural `Flair`-based approach outperforms standard SVM-based methods. In the work of Basaldella *et al.* [37], the authors pretrained `ELMo` and `Flair` contextualized word embeddings on health forums within Reddit and applied them to health social media data for various NER problems. They concluded that domain-based contextualized word embeddings heavily influence the performance on downstream tasks, outperforming embeddings trained both on general-purpose data or on scientific papers when applied to user-generated content. Our experiments are very similar to this work.

One can find an extensive overview of recent advances in NLP field in the work of Minaee *et al.* [38]. While focusing on document classification, it describes several methods, such as transformers, which are completely applicable to NER. Young *et al.* [39] cover in detail a key element of the current paper – distributed and contextualized word representation, among other recent trends introduced in NLP.

B. SPANISH CLINICAL TEXT PROCESSING

Spanish is an inflectional language with a richer morphology compared to the English language; morphemes denote several syntactic, semantic and grammatical features of words (such as gender, number, etc). From a syntactic point of view, Spanish texts have more subordinate clauses and long sentences with a high word order flexibility; for instance, the subject can be located in any position in a sentence instead of only before the verb.

There are a number of peculiarities of clinical texts in Spanish. Due to translation of English biomedical terms, there are more variants of anglicisms. Some of them are freely adapted and others are exact copies of original ones, for instance ''interleukin'' is translated to ''interleukina''/''interleucina''/''interleuquina''. Moreover, Spanish language uses accent marks which do not exist in English and the preference or not of using these generates lexical variants; for instance, ''period'' may be transformed into ''period'' or ''período''. Adjectives ending in ''-al'' sometimes keep their form when translated and sometimes follow Spanish morphological rules, for example, ''viral'' may be transformed to ''viral'' or ''vírico'' and ''bacterial'' to ''bacterial''/''bacteriano''/''bacteriana''/''bacterianos''/''bacterianas'' (considering gender and number morphological variants).

Greco-latin prefixes show variants like ''psi-'' (''psicologo'' vs ''sicólogo'') or ''pseudo-''. The use of hyphens between words is more systematic in English while in Spanish many variants occur. For instance, ''beta-carotene'' is transformed into ''beta-caroteno''/''beta caroteno''/''betacaroteno''/''caroteno beta''. The names of pharmacological substances sometimes remain the same as in English and others are adapted, e.g. ''furazosin'' is adapted to ''furazosina''/''furazosín''/''furazosin''. Concerning gender (male/female), in some terms there is ambiguity (''la COVID''/''el COVID'') or both are allowed, for instance, ''el tiroides'' (male) /''la tiroides'' (female) for ''thyroid'' hormone.

Clinical notes have many occurrences of abbreviations and usually English abbreviations coexist with Spanish ones. For instance, ''PSA'' corresponds to ''prostate-specific antigen'' and it is preferred to ''APE'' (''antígeno prostático específico''). However, polysemic abbreviations are very common in both languages.

From a syntactic point of view, sentences are very similar in both languages (short sentences or phrases, with use of negation particles and non-standard abbreviations, misspellings, speculation and ungrammatical sentences, among other phenomena).

In summary, there are more lexical variants of medical terms in Spanish with respect to English due to replication or partial adaptation of terms. For these reasons, analyzing these texts is a more resource consuming task, and normalization tools are required.

C. PharmacoNER 2019 SHARED TASK

PharmacoNER is ''the first task on chemical and drug mention recognition from Spanish medical texts, namely from a corpus of Spanish clinical case studies'' [1]. According to the organizers, ''the main aim was to promote the development of named entity recognition tools of practical relevance, that is, chemical and drug mentions in non-English content, determining the current-state-of-the art, identifying challenges and

comparing the strategies and results to those published for English data''.

The challenge consisted of two subtracks – (1) NER offset and entity classification and (2) Entity indexing. We focus on the NER task. In total, 22 teams participated in the first subtrack. Xiong *et al.* [40] was placed first with an overall F-score of 91.05. They used the multi-lingual large version of the pre-trained BERT model[4] with further fine-tuning to the PharmacoNER NER problem. The key success of their implementation of the BERT model in comparison to other participants' BERT implementations was that they incorporated more semantic and syntactic features such as word shape and PoS tags into their model embedding layer. Moreover, they applied a Spanish biomedical abbreviation detection tool, however they did not detail how the extracted abbreviations were further used.

The second-best results of Stoeckel *et al.* [41] were updated after the formal challenge with an F-score of 90.52. They used the Flair model and made use of additional corpus derived from SciELO, however of a smaller size than ours. They used this corpus to train word2vec and FastText word embeddings, and for Flair language model (LM)-based embeddings they used pre-trained Spanish general domain word embeddings.[5]

Sun *et al.* [42] achieved the third-best result with an F-score of 89.24. Like Xiong *et al.* [40], they also used the pre-trained version of BERT with subsequent fine-tuning but without incorporating any additional features.

Overall, many participants experimented with document encoding techniques. For example, Rivera Zavala *et al.* [43] gathered similar size Spanish biomedical corpora to train their own FastText embeddings. Moreover, they used sense2vec [44] pre-trained embeddings. Both of these embeddings have proven useful in extracting biomedical concepts.

Later, other research papers appeared addressing NER on the PharmacoNER corpus. Multi-tasked and stacked model approaches were offered by [45]. Their best multi-tasking approach achieved 91.4 F-score. In another paper, a set of 104 sophisticated context patterns was constructed [46]. With this knowledge-based approach, authors achieved an impressive result of 91 F-score. We do not compare our results with the results of these two papers, as the approach of [45] required more annotated data, and the approach of [46] required manual rule construction relying on Spanish language syntax.

II. METHODS

A. FLAIR

Flair embeddings were developed by the Zalando research group [17]. They are contextualized string embeddings in the sense that the contextualized embedding vectors are trained without any notion of words but purely treat texts as sequences of characters. This is the main difference between this type of embeddings and others such as word2vec [47], Glove [7], and ELMo [17].

Flair is trained using an LM objective function aimed at predicting the next character of a sequence, thus keeping information on the character ordering in a text sequence. By learning the character level representations in both directions it was possible to get the context for each character in both right and left directions. To generate a word embedding from characters the first and last character states of each word are extracted and concatenated.

From the computational and memory point of view, these embeddings are more efficient to store and train a model for word embeddings. Moreover, they have proven to be more effective in terms of rare, out-of-vocabulary (OOV) words and morphologically rich languages [17].

In our experiments, we use the enhanced version of Flair embeddings called Pooled Contextualized String Embeddings [48]. It is different from the previously developed Flair LM in that it better handles representation for words in an underspecified context. By dynamically aggregating the contextualized embedding of each unique word, this information is later used to expand the embedding for the same word encountered in a poorly, ambiguously specified context. This situation is often encountered in the Spanish biomedical NER tasks, when two words with similar suffixes express different types of substances, as for example, *creatinina* and *hemoglobina* where the latter is a protein but the former is not.

B. BERT

Bidirectional Encoder Representations for Transformers (BERT) is the deep learning language representation model developed by the Google research team [11]. In contrast to ELMo and Flair, it can be used not only for contextualized word embeddings generation, but also for the downstream tasks itself through a process called *fine-tuning*.

BERT is trained using the masked word piece representation and the next sentence objective. Its architecture consists of stacked multi-layered transformers, each having a self-attention mechanism with multiple attention heads. Introducing self-attention in encoder-decoder architecture of BERT allows better capture of long-distance relationships among concepts by avoiding no locality bias.

BERT can be further pre-trained for a specific domain or fine-tuned for a specific task [49]. In particular, fine-tuning for token level classification tasks is supported by putting a linear layer, which takes as an input the last hidden state of the sequence, on top of the BERT model.

C. ADDITIONAL EMBEDDINGS

It has been demonstrated that the concatenation of contextualized embeddings with the standard embeddings usually leads to an improvement in results [10], [17]. Following this, for our experiments we used the concatenation of Flair embeddings with Spanish general

[4]https://storage.googleapis.com/bert_models/2018_11_03/multilingual_L-12_H-768_A-12.zip

[5]http://www.github.com/iamyihwa

(not domain specific) FastText embeddings [8], domain-specific Spanish biomedical FastText embeddings [18], byte-pairwise encoded embeddings (BPE) [50] and character embeddings [51]. The results of models with and without these additional embeddings are presented.

General FastText embeddings for Spanish were trained using the full dump of Spanish-language Wikipedia while Spanish domain-specific biomedical embeddings utilizing the architecture of FastText were trained over the SciELO[6] corpus with 100 million tokens and the health section of Wikipedia with 82 million tokens.

Character embeddings are generated using a RNN model and further are concatenated with the other types of word embeddings in a model.

While the BPE model represents subword embeddings in 275 languages, we used only one language from this model. It produces relatively light-weight embeddings as they consist of sub-word tokens of words. This method has been shown to deal well with unknown words and to produce results on a par with the standard word embeddings.

D. ENTITY EXTRACTION

In the PharmacoNER task, there are 4 relevant types of entity mentions, although for the official evaluation, only the first 3 types are used:

- **Normalizables** (Normalizable): mentions of biomedical concepts which can be normalized to the SNOMED-CT and ChEBI vocabularies;
- **No_Normalizables** (Non-normalizable): biomedical concepts which cannot be normalized to the given vocabularies;
- **Proteínas** (Proteins): mentions of genes and proteins;
- **Unclear** (Unclear): general substance mentions.

The problem of biomedical NER can be framed as a sequence labeling task where the goal is to extract the correct spans of entities. We therefore used a BIO schema. In this schema, each token in a document is classified as [B]eginning, [I]nside, or [O]utside of an entity mention.

Other than for the BERT experiments, all experiments were conducted using the Flair framework[7] which is built on top of Theano providing a convenient means of experimenting with different combinations of word embeddings. It provides an off-the-shelf neural-based system supporting entity extraction. We train a Long Short Term Memory (LSTM) network with a hidden state of 256 dimensions, learning rate 0.1, mini-batch size of 8, and is optimized with Adam. We train for 150 epochs, and the model that performs best on the validation set provided by the organizers of the competition during training is used to prevent overfitting.

We were unable to conveniently experiment with BERT embeddings using the Flair framework but preferred the Google Cloud TensorFlow TPU set up for both training contextualized word embeddings and the downstream task fine-tuning and predictions as it works much faster.[8] However, at the time of writing TPU did not support inference on downstream tasks, and it was required to switch over to CPU instances for this step.

We used a Conditional Random Fields loss [52] as it has been shown to increase the accuracy for the NER tasks. The training and evaluation batch sizes were set to 32 and 8, respectively, and the learning rate was set to $5e^{-5}$. The maximum sequence length was set to 160.

Despite the common advice to fine-tune the BERT model for just 3-10 epochs, we fine-tuned it for 30 epochs as we noticed it improved the predictions.

E. PharmacoNER CORPUS

The PharmacoNER corpus was used for training and testing our models. It consists of 1000 annotated SPACCC articles derived from open access Spanish medical publications in SciELO – an electronic library where complete full-text articles from scientific journals of Latin America, South Africa, and Spain are systematically collected and stored.[9]

TABLE 1. Statistics on PharmacoNER corpus.

Size (sent)	Size (words)	Entity types and counts
16,504	396988	Normalizables (4,426),
16.5 sent/case	396.2 words/case	No_Normalizables (55),
		Proteinas (2,291), Unclear (159)

Table 1 shows summary statistics of the PharmacoNER corpus. Results are scored with the scoring tool distributed by the organizers of the challenge. For concepts, true positives are strict (the system concept span must match a gold concept spans begin and end exactly). We report micro-averaged results of the lenient evaluation since that was the metric used to score the shared task.

For training the model, we combined both training and development corpora (yielding 11970 sentences for the merged corpora) and selected by random shuffling 10% of it for validation purposes.

F. LANGUAGE MODEL TRAINING DATASET

We selected a subset of SciELO text based on some heuristics to be in line with the corpus used for training and testing the model. In particular, we chose articles based on the criteria that the specified area of the document is Health Sciences and then selected text in particular sections of the articles. Specifically, text starting with section headings 'Descripción del caso', 'Presentación de caso', 'Descripción de caso clínico', or 'Caso clínico', and ending with the sections 'Bibliografía' or 'Referencias' was selected. In this way we retrieved 1, 368, 080 sentences with 86, 851, 275 tokens.

We also used the same corpora for training the BERT language representation model with the vocabulary size set to 128000.

[6] SciELO.org
[7] https://github.com/zalandoresearch/flair
[8] https://cloud.google.com/ml-engine/docs/tensorflow/using-tpus
[9] http://www.scielo.org

TABLE 2. Results of experiments.

	Sun's BERT			Stoeckel's Flair			Xiong's BERT			Flair_Sc_ext2 (ours)			BERT_Sc (ours)		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Overall	90.46%	88.06%	89.24%	90.79%	90.30%	90.52%	91.23%	**90.88%**	**91.05%**	**91.97%**	89.74%	90.84%	89.29%	87.83%	88.55%
Normalizables	-	-	-	-	-	-	94.26%	**92.91%**	**93.58%**	**95.21%**	91.88%	93.46	91.48%	91.67%	91.57%
Proteinas	-	-	-	-	-	-	87.87%	**89.41%**	**88.63%**	**88.56%**	88.36%	88.46%	86.74%	84.52%	85.61%
No Normalizables	-	-	-	-	-	-	100.00%	20.00%	**33.33%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Constructed domain-specific Flair embeddings Flair_Sc are compared with the general pre-trained Flair embeddings that are a part of the Flair API Flair_G.[10] They are trained on a dump of Spanish Wikipedia dated August 2018.[11]

G. LANGUAGE MODEL TRAINING

The Flair_Sc LM was trained until the perplexity reached 1.92. The settings used to train word embeddings are: hidden size 1150, the number of layers 3 with maximum sequence length 240, mini-batch size 100 and number of epochs equal to 1000.

The training of Flair_Sc LM was done using 1 GPU instance.

The BERT language representation was trained using Tensor Processing Units (TPU) instances in Google Colab with the number of training steps 1B. TPU is designed to efficiently scale operations among different machines thus making calculations on tensors faster than doing it using GPU. For storing and uploading weights for training Google Cloud persistent storage is required. Moreover, every 8 hours Google Colab shuts down its server, so it is necessary to be resumed manually. Overall, it took more than 4 days to train the BERT language representation, substantially longer than it takes to train Flair_Sc LM.

III. RESULTS

The comparative results of experiments are presented in Table 2 where we depict our best Flair-based, and BERT model results:

- **Flair_Sc_ext2**: the extended model is trained using the custom SciELO Flair embeddings Flair_Sc, SciELO FastText embeddings, BPE embeddings and character Embeddings;
- **BERT_Sc**: BERT-based word embeddings are trained on the SciELO corpus. Subsequently, the BERT model is fine-tuned for the downstream task.

To compare our results with others, we selected the top results in the challenge leader board and we omit results for which no description of the systems were provided.

For the best model, precision for all types of entities is higher than recall, especially for *Normalizables* entities. This means that while the model is good in determining the correct cases, it is not as strong at identifying positive examples.

Indeed, comparing to the best systems' results, it can be observed that we are superior in terms of higher precision but relatively weaker in terms of recall. Overall, our results are 0.21 points behind the best system of Xiong *et al.* [40] for this task.

No_Normalizables entities comprising the minority class are not captured by our models. Techniques for tackling the class imbalance should be considered in future experiments with sequence labeling architectures.

IV. DISCUSSION

A. NUMBER OF TRAINING EPOCHS

Fig. 1 shows an evolution of the loss and F-score over number of epochs. It can be seen that the loss becomes steady after around 27 epochs and the test F-score stabilizes at around the same point. Overall, the test set loss curve resembles the validation set loss curve which means that the validation set is a good proxy for measuring the model performance.

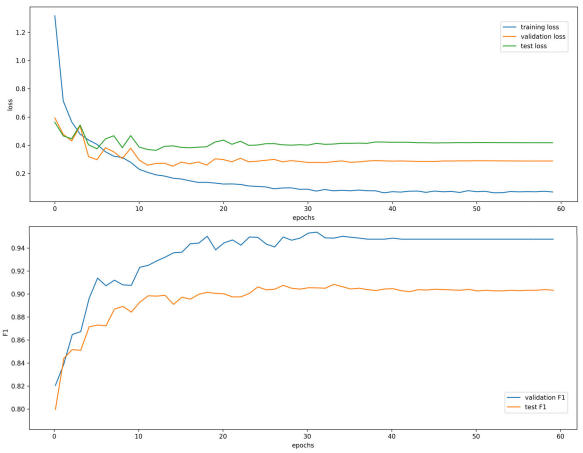

FIGURE 1. Training and validation loss and F-score Dependency between the number of epochs and either loss (top figure) or F1 score (bottom figure) for Flair_Sc_ext2 model.

B. ABLATION ANALYSIS

For our ablation analysis, we explored the following additional combinations of word embeddings:

- **Flair_G_ext**: the model is trained using Spanish general domain Flair embeddings Flair_G, Spanish general FastText embeddings and BPE embeddings;
- **Standard_Sc**: SciELO FastText embeddings with subword information property, BPE embeddings and character Embeddings are used;
- **Flair_Sc**: based only on custom SciELO Flair embeddings;
- **Flair_Sc_ext**: the custom SciELO Flair embeddings, general Spanish FastText embeddings and BPE embeddings are used.

[10] https://github.com/zalandoresearch/flair/issues/80
[11] https://dumps.wikimedia.org/eswiki/20180801/

TABLE 3. Results of ablation analysis.

Model name	Embedding types	Precision	Recall	F-score
`Flair_G_ext`	General `Flair` + general `FastText` + `BPE`	89.71%	89.47%	89.59%
`Standard_Sc`	SciELO `FastText` + `BPE` + char emb	86.90%	86.81%	86.85%
`Flair_Sc`	SciELO `Flair`	88.91%	88.38%	88.65%
`Flair_Sc_ext`	SciELO `Flair` +general `FastText` + `BPE`	90.95%	89.47%	90.20%
`Flair_Sc_ext2`	SciELO `Flair` + SciELO `FastText` + `BPE` + char emb	**91.97%**	**89.74%**	**90.84%**

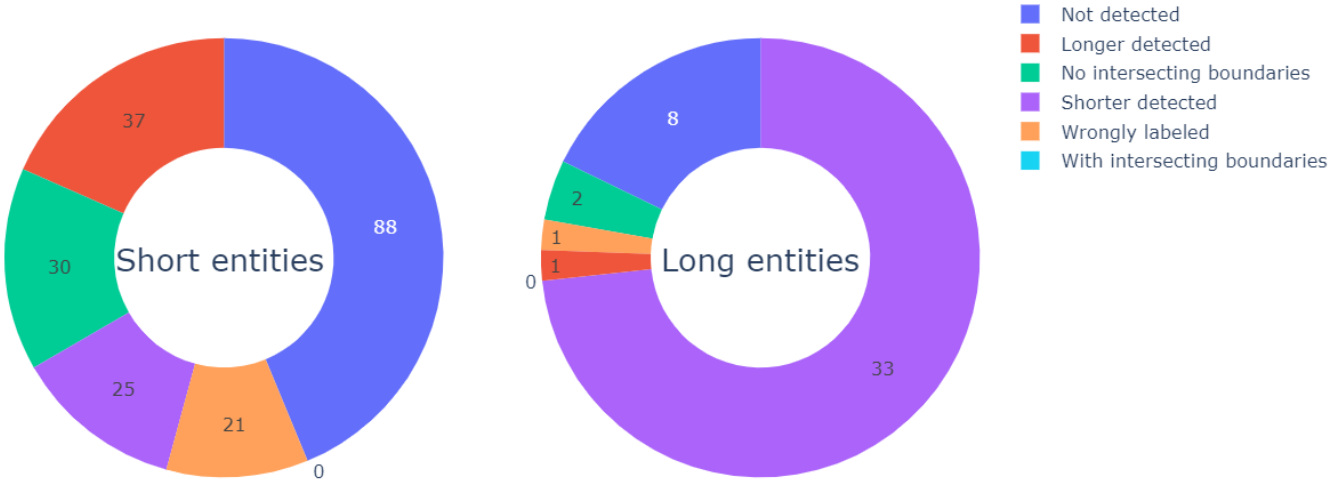

FIGURE 2. Distribution of errors Distribution of errors for short (less than 3 terms) and long (with length more or equal 3 terms) entities.

The results of different variations of stacking word embeddings are shown in Table 3.

In general, LM-based embeddings lead to better results than the standard ones. It can be also seen that the model enriched with different types of word embeddings gives better results in terms of precision, recall and F-score. Domain specific word embeddings lead to improvement of results, however, they are much smaller in size than general domain ones. Augmenting word embeddings with additional subword level embeddings such as `FastText`, `BPE` and character embeddings further improves the results.

We also experimented with searching concepts in SNOMED-CT using the Meaning Cloud tool,[12] however it did not work well, as many concepts for the shared task were annotated based on their synonyms.

C. ERROR ANALYSIS

For error analysis, we split gold standard entities into 2 groups: short entities with the length less or equal to 2, and long entities with the length greater than or equal to 3. For the best model `Flair_Sc_ext2`, the origin and distribution of errors are presented in Fig. 2.

It can be seen that the majority of errors are for the short predicted entities for which there is not even partial overlap with gold standard entities (No intersections false positives (FP)). Indeed, many biomedical entities are acronyms and abbreviations which could be easily misclassified based on casing and length of entities. Interestingly, the second primary source of errors for short predicted entities is that the model predicts two entities where the gold standard has a single entity (Longer FP). A smaller number of errors are related to short gold standard entities which the model fails to detect (false negatives - FN). For long entities, the main source of error is that predicted entities are shorter than required (Shorter FP), contributing nearly 75% of the total error.

In Table 4, we present a comparison of errors among 3 models: the best model `Flair_Sc_ext2`, the model which uses only `Flair` embeddings trained over the target corpus `Flair_Sc`, and the model based on a set of standard embeddings `Standard_Sc`.

Interestingly, the main discrepancies in the number of errors for `Flair_Sc` model in comparison to the best model are related to the larger number of not predicted short entities (FN). All other discrepancies in errors for both models vary in a range 1-7 in both ways.

In relation to the best model, the main source of errors for `Standard_Sc` is related as well to the falsely predicted short entities without intersections with gold standard ones (No intersections FP) with almost 15% more predicted FP. It indicates that the best model utilizing the contextual embeddings learns the meaning of acronyms, abbreviations and overall short uppercased words more effectively, assigning them biomedical labels with more caution.

This comparison also shows that lower performing models are much worse at detecting the boundaries of short biomedical concepts, often predicting longer concepts:

[12]https://www.meaningcloud.com/

TABLE 4. Distribution of errors for different models.

	Short ($\leq$ 2 terms)			Long ($\geq$ 3 terms)		
	Flair_Sc_ext2	Flair_Sc	Standard_Sc	Flair_Sc_ext2	Flair_Sc	Standard_Sc
Misclassified FP	21	28	30	1	0	2
Shorter FP	25	27	22	33	38	31
Longer FP	37	42	58	1	0	1
Intersections FP	0	2	3	0	1	0
No intersections FP	88	84	103	8	4	6
FN	30	50	72	2	1	1

TABLE 5. Examples of errors in recognizing biomedical entities by different models.

		Example	Types of errors
Example 1	Correct annotation	IgG 317, IgA 1446, IgM 15 mg/dl, cadena ligera libre (CLL, nefelometría Free-Lite®) kappa 4090 ng/ml, lambda 1.	
	Flair_Sc_ext2	IgG 317, IgA 1446, IgM 15 mg/dl, cadena ligera libre (CLL, nefelometría Free-Lite®) kappa 4090 ng/ml, lambda 1.	FP
	Flair_Sc	IgG 317, IgA 1446, IgM 15 mg/dl, cadena ligera libre (CLL, nefelometría Free-Lite®) kappa 4090 ng/ml, lambda 1.	FP
	Standard_Sc	IgG 317, IgA 1446, IgM 15 mg/dl, cadena ligera libre (CLL, nefelometría Free-Lite®) kappa 4090 ng/ml, lambda 1.	FP
Example 2	Correct annotation	proteína S-100 (Dako, L1845, USA, prediluída), neurofilamentos (Biogenex 6670-0154, USA), enolasa neuroespecífica NSE,	
	Flair_Sc_ext2	proteína S-100 (Dako, L1845, USA, prediluída), neurofilamentos (Biogenex 6670-0154, USA), enolasa neuroespecífica NSE,	FP, shorter FP
	Flair_Sc	proteína S-100 (Dako, L1845, USA, prediluída), neurofilamentos (Biogenex 6670-0154, USA), enolasa neuroespecífica NSE,	FP, shorter FP
	Standard_Sc	proteína S-100 (Dako, L1845, USA, prediluída), neurofilamentos (Biogenex 6670-0154, USA), enolasa neuroespecífica NSE,	FP, longer FP

5 more incorrectly predicted concepts for the Flair_Sc model and 21 more incorrectly predicted concepts for the Standard_Sc model.

It is interesting to observe that for the long predicted concepts, the absolute numbers and distribution of errors for the best Flair_Sc_ext2 and Standard_Sc models are mostly the same. However, the Flair_Sc model performs slightly worse in terms of predicting shorter concepts than the gold standard ones (i.e. predicting three consecutive terms instead of four, etc).

In Table 5, we present two examples of sentences with underlined gold standard and predicted entities. Sentences were chosen from the representative groups of the most common errors for different models. Here, *FP* is the shorter abbreviation for FP without intersections. It can be observed that the Standard_Sc model in both examples predicted long entities which were either FP or longer version of gold standard entities. Flair-based models are also often confusing short upper-cased entities but in fewer cases.

Interestingly, in the second example, although both Flair_Sc and Standard_Sc models have detected 'USA' entity as a PROTEIN, the Flair_Sc_ext2 model which combines embeddings from both models did not give this entity a biomedical label.

In terms of the best parameter setting, we did not perform parameter selection for either the Flair or BERT models; this might further increase model quality.

V. CONCLUSION

In this work, we have explored the application of transfer learning techniques, in particular, language representation-based word embeddings to the problem of extracting biomedical entities from 1000 Spanish clinical case narratives. By leveraging the knowledge from a huge amount of unlabeled data, with language model pre-training it becomes possible to build a high-quality NER system even with this small amount of annotated data.

With this aim, we trained domain-specific Spanish language models, in particular, Flair and BERT to derive contextualized word embeddings and applied them to the PharmacoNER biomedical NER data achieving competitive results. We showed that domain-specific word embeddings outperform general embeddings, despite being trained on a smaller corpus. Moreover, we demonstrated that stacking together word embeddings of different nature can improve model performance.

Error analysis has shown that the main source of errors for all models is over-zealous recognition of short entities. Indeed, biomedical entities are often short and upper-cased and can be easily mixed up with other abbreviated short words. Testing the approach by analyzing other Spanish health-related texts, such as social media [53], with similar characteristics (e.g., a large number of abbreviations, lack of grammatical structure, punctuation marks, etc.) and others (e.g., patient oriented terminology not included in any resource, slang words, etc.) could help to cope with these phenomena.

Moreover, standard embedding-based models often fail by detecting long false positive entities or longer versions of gold standard entities (in particular, for FastText models). However, it should be noted that the ability to detect long entities could be beneficial in particular scenarios.

One direction for improvement could be more sophisticated utilization of contextualized embeddings. For example, they could be incorporated into state-of-the-art NER architectures such as graph-based NNs or NNs with a dependency tree-based attention mechanism to further improve capturing of long-distance relationships between biomedical entities.

For handling the imbalance of classes, different strategies such as loss function modification could be applied in future work.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their invaluable feedback. They are also thankful to L. Campillos who gratefully helped us prepare the analysis on Spanish clinical texts' peculiarities.

REFERENCES

[1] A. G. Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, and M. Krallinger, "PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track," in *Proc. 5th Workshop BioNLP Open Shared Tasks*, 2019, pp. 1–10. [Online]. Available: https://github.com/PlanTL-SANIDAD/SPACCC

[2] I. Segura-Bedmar, R. Revert, and P. Martínez, "Detecting drugs and adverse events from Spanish social media streams," in *Proc. 5th Int. Workshop Health Text Mining Inf. Anal. (Louhi)*, 2014, pp. 106–115.

[3] L. Akhtyamova, A. Ignatov, and J. Cardiff, "A large-scale CNN ensemble for medication safety analysis," in *Natural Language Processing and Information Systems. NLDB* (Lecture Notes in Computer Science), vol. 10260, F. Frasincar, A. Ittoo, L. Nguyen, and E. Métais, Eds. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-59569-6_29.

[4] D. Kim, J. Lee, C. H. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung, and J. Kang, "A neural named entity recognition and multi-type normalization tool for biomedical text mining," *IEEE Access*, vol. 7, pp. 73729–73740, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8730332/

[5] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 2567–2577. [Online]. Available: http://arxiv.org/abs/1909.06146

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[7] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: https://nlp.stanford.edu/pubs/glove.pdf

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 7, no. 5, pp. 135–146, 2017. [Online]. Available: http://arxiv.org/abs/1607.04606

[9] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017. [Online]. Available: https://academic.oup.com/bioinformatics/article/33/14/i37/3953940

[10] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[12] X. Han and J. Eisenstein, "Unsupervised domain adaptation of contextualized embeddings for sequence labeling," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4237–4247. [Online]. Available: https://github.com/xhan77/

[13] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 328–339. [Online]. Available: http://nlp.fast.ai/ulmfit

[14] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 3079–3087. [Online]. Available: http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning.pdf

[15] A. Radford. (2018). *Improving Language Understanding by Generative Pre-Training*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf and https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: http://arxiv.org/abs/1706.03762

[17] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proc. COLING*, 2018, pp. 1638–1649. [Online]. Available: https://github.com/zalandoresearch/flair

[18] F. Soares, M. Villegas, A. Gonzalez-Agirre, M. Krallinger, and J. Armengol-Estapé, "Medical word embeddings for Spanish: Development and evaluation," in *Proc. 2nd Clin. Natural Lang. Process. Workshop*, 2019, pp. 124–133, doi: 10.5281/zenodo.2542722.

[19] L. Akhtyamova, "Named entity recognition in Spanish biomedical literature: Short review and bert model," in *Proc. 26th Conf. Open Innov. Assoc. (FRUCT)*, Apr. 2020, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/9087359/

[20] T. Eftimov, B. K. Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0179488, doi: 10.1371/journal.pone.0179488.

[21] C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter, and K. Verspoor, "Large-scale biomedical concept recognition: An evaluation of current automatic annotators and their parameters," *BMC Bioinf.*, vol. 15, no. 1, p. 2, Dec. 2014.

[22] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, and M. Krallinger, "Automatic de-identification of medical texts in Spanish: The MEDDOCAN track, corpus, guidelines, methods and evaluation of results," in *Proc. Iberian Lang. Eval. Forum (IberLEF)*, 2019, pp. 618–638. [Online]. Available: https://github.com/PlanTL-SANIDAD

[23] N. Perez, L. García-Sardiña, M. Serras, and A. D. Pozo, "Vicomtech at MEDDOCAN: Medical document anonymization," in *Proc. Iberian Lang. Eval. Forum (IberLEF)*, 2019, pp. 698–703. [Online]. Available: https://github.com/PlanTL-SANIDAD/SPACCC

[24] P. López-Ubeda, M. C. Díaz-Galiano, L. A. U. López, and M. Teresa, "Anonymization of clinical reports in Spanish: A hybrid method based on machine learning and rules," in *Proc. Iberian Lang. Eval. Forum (IberLEF)*, 2019, pp. 688–695. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2018T01

[25] K.-J. Lee, Y.-S. Hwang, S. Kim, and H.-C. Rim, "Biomedical named entity recognition using two-phase model based on SVMs," *J. Biomed. Inform.*, vol. 37, no. 6, pp. 436–447, Dec. 2004.

[26] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus—A semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. 1, pp. i180–i182, Jul. 2003. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btg1023.

[27] T. V. T. Nguyen, A. Moschitti, and G. Riccardi, "Kernel-based reranking for named-entity extraction," in *Proc. 23rd Int. Conf. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 901–909. [Online]. Available: https://dl.acm.org/citation.cfm?id=1944670

[28] M. Collins, "Ranking algorithms for named-entity extraction," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001, pp. 489–496.

[29] J. Björne and T. Salakoski, "Generalizing biomedical event extraction," in *Proc. BioNLP Shared Task Workshop*, 2011, pp. 183–191. [Online]. Available: http://svmlight.joachims.org/svm_

[30] K. Takeuchi and N. Collier, "Bio-medical entity extraction using support vector machines," *Artif. Intell. Med.*, vol. 33, no. 2, pp. 125–137, Feb. 2005.

[31] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," in *Proc. 19th Int. Conf. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 1–7.

[32] R. Patra and S. K. Saha, "A kernel-based approach for biomedical named entity recognition," *Sci. World J.*, vol. 2013, pp. 1–7, Jan. 2013.

[33] D. Li, L. Huang, H. Ji, and J. Han, "Biomedical event extraction based on knowledge-driven tree-LSTM," in *Proc. NAACL-HLT*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 1421–1430.

[34] L. Simeonova, K. Simov, P. Osenova, and P. Nakov, "A morpho-syntactically informed LSTM-CRF model for named entity recognition," 2019, *arXiv:1908.10261*. [Online]. Available: http://arxiv.org/abs/1908.10261 and http://github.com/lilia-simeonova/

[35] S. Sharma and R. Daniel, Jr., "BioFLAIR: Pretrained pooled contextualized embeddings for biomedical sequence labeling tasks," 2019, *arXiv:1908.05760*. [Online]. Available: http://arxiv.org/abs/1908.05760
[36] T. Miller, A. Geva, and D. Dligach, "Extracting adverse drug event information with minimal engineering," in *Proc. 2nd Clin. Natural Lang. Process. Workshop*, 2019, pp. 22–27. [Online]. Available: https://www.aclweb.org/anthology/W19-1903
[37] M. Basaldella and N. Collier, "BioReddit: Word embeddings for user-generated biomedical NLP," in *Proc. 10th Int. Workshop Health Text Mining Inf. Anal. (LOUHI)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 34–38.
[38] S. Minaee, E. Cambria, and J. Gao, "Deep learning based text classification: A comprehensive review," 2020, *arXiv:200403705*. [Online]. Available: https://arxiv.org/abs/2004.03705
[39] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [Review Article]," Inst. Elect. Electron. Eng. Inc., New York, NY, USA, Tech. Rep., 2018.
[40] Y. Xiong, Y. Shen, Y. Huang, S. Chen, B. Tang, X. Wang, Q. Chen, J. Yan, and Y. Zhou, "A deep learning-based system for PharmaCoNER," in *Proc. 5th Workshop BioNLP Open Shared Tasks*, 2019, pp. 33–37. [Online]. Available: https://github.com/PlanTL-SANIDAD/SPACCC_POS-
[41] M. Stoeckel, W. Hemati, and A. Mehler, "When specialization helps: Using pooled contextualized embeddings to detect chemical and biomedical entities in Spanish," in *Proc. 5th Workshop BioNLP Open Shared Tasks*, 2019, pp. 11–15. [Online]. Available: https://www.github.com/zalandoresearch/flair
[42] C. Sun and Z. Yang, "Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task," in *Proc. 5th Workshop BioNLP Open Shared Tasks*, 2019, pp. 100–104.
[43] R. Rivera and P. Martínez, "Deep neural model with enhanced embeddings for pharmaceutical and chemical entities recognition in spanish clinical text," in *Proc. 5th Workshop BioNLP Open Shared Tasks*, 2019, pp. 38–46. [Online]. Available: https://ufal.mff
[44] A. Trask, P. Michalak, and J. Liu, "Sense2vec—A fast and accurate method for word sense disambiguation in neural word embeddings," 2015, *arXiv:1511.06388*. [Online]. Available: http://arxiv.org/abs/1511.06388
[45] L. Lange, H. Adel, and J. Strötgen, "Closing the gap: Joint de-identification and concept extraction in the clinical domain," 2020, *arXiv:2005.09397*. [Online]. Available: http://arxiv.org/abs/2005.09397
[46] F. S. León and A. G. Ledesma, "Annotating and normalizing biomedical NEs with limited knowledge," 2019, *arXiv:1912.09152*. [Online]. Available: http://arxiv.org/abs/1912.09152
[47] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119. [Online]. Available: https://arxiv.org/pdf/1310.4546.pdf
[48] A. Akbik, T. Bergmann, and R. Vollgraf, "Pooled contextualized embeddings for named entity recognition," in *Proc. NAACL*, 2019, pp. 724–728. [Online]. Available: https://github.com/zalandoresearch/flair
[49] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2019. [Online]. Available: https://github.com/dmis-lab/biobert
[50] B. Heinzerling and M. Strube, "BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1418–1473. [Online]. Available: https://aclweb.org/anthology/papers/L/L18/L18-1473/
[51] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 260–270. [Online]. Available: http://aclweb.org/anthology/N16-1030
[52] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*. San Francisco, CA, USA: Morgan Kaufmann, 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813
[53] P. Martínez, J. L. Martínez, I. Segura-Bedmar, J. Moreno-Schneider, A. Luna, and R. Revert, "Turning user generated health-related content into actionable knowledge through text analytics services," *Comput. Ind.*, vol. 78, pp. 43–56, May 2016.

LILIYA AKHTYAMOVA received the M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, State University, Dolgoprudniy, Moscow obl., Russia, in 2017. She is currently pursuing the Ph.D. degree in information technology at the Technological University Dublin, Dublin, Ireland.

In Spring 2019, she did the traineeship under the Erasmus grant at the Carlos III University of Madrid, Madrid, Spain. This article resulted from the collaboration work performed at that time. Her research interests include biomedical text mining, named entity recognition, and deep learning.

PALOMA MARTÍNEZ received the degree in computer science and the Ph.D. degree in computer science from the Universidad Politécnica de Madrid, Spain, in 1992 and 1998, respectively.

She is the Head of the Human Language and Accessibility Technologies (HULAT) in the Computer Science and Engineering Department, University Carlos III of Madrid. Her research interests are human language technologies, with the focus on information extraction in the biomedical domain, and web accessibility. She is the coauthor of more than 40 articles in indexed journals and more than a hundred international conference contributions. She has been principal investigator and participated in over 40 national and international research projects. She is currently a member of the Spanish Society for Natural Language Processing (SEPLN) and a member of Dynamization Network for Activities on Natural Language Processing Technologies. She is a collaborator of the Spanish Center of Captioning and Audiodescription (CESyA).

KARIN VERSPOOR received the B.A. degree in computer science and cognitive sciences from Rice University, in 1993, and the M.Sc. degree in cognitive science and natural language and the Ph.D. degree in cognitive science from the University of Edinburgh, U.K., in 1994 and 1997, respectively.

After a post-doc at Macquarie University in Sydney, Australia, she spent five years in artificial intelligence start-ups, and then held research roles at the Los Alamos National Laboratory, the University of Colorado School of Medicine, National Information Communications Technology Australia (NICTA), and finally joined the University of Melbourne. This work was completed while she was a long-term Visitor at The University of Carlos III Madrid, Spain, hosted by Paloma Martínez and with the support of the University of Melbourne. Her research focuses on biomedical text mining and clinical data analysis.

JOHN CARDIFF received the B.A. degree (Hons.) in computer science from the Trinity College Dublin, Ireland, in 1986, and the Ph.D. degree from the University of Queensland, Australia, in 1990.

He has over 25 years lecturing and research experience and is currently a full time Lecturer at the Technological University of Dublin (Tallaght Campus), Ireland. He has previously held positions in the Department of Computer Science, Trinity College Dublin, and in the University of Queensland, Australia. He has served as Visiting Professor at the Technical University of Valencia, Spain and Universitat Jaume I, Spain. His research interests include Natural Language Processing and Social Media Analysis. He is author or co-author of over 80 scientific papers.