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

Liliya Akhtyamova (akhtyamova@phystech.edu)
Technological University Dublin https://orcid.org/0000-0003-4338-1483
Paloma Martínez
Universidad Carlos III de Madrid
Karin Verspoor
University of Melbourne
John Cardiff
Technological University Dublin

Technical advance

Keywords: Natural language processing, Named entity recognition, Clinical case narratives, Deep learning, Language representations, Contextualized word embeddings, Spanish language

Posted Date: February 5th, 2020

DOI: https://doi.org/10.21203/rs.2.22697/v1

Version of Record: A version of this preprint was published on August 24th, 2020. See the published version at https://doi.org/10.1109/ACCESS.2020.3018688.

BMC Medical Informatics and Decision Making

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

--Manuscript Draft--

Manuscript Number:	MIDM-D-20-00058
Full Title:	Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives
Article Type:	Technical advance
Section/Category:	Standards, technology, machine learning, and modeling
Funding Information:	Ministerio de Economía y Competitividad (TIN2017-87548-C2-1-R) — Ms Paloma Martínez
Abstract:	Background: In the Big Data era there is an increasing need to fully exploit and analyse the huge quantity of information available about health. Natural Language Processing (NLP) technologies can contribute to extract relevant information from unstructured data contained in Electronic Health Records (EHR) such as clinical notes, patient's discharge summaries and radiology reports among others. Extracted information could help in health-related decision making processes. Named entity recognition (NER) devoted to detect important concepts in texts (diseases, symptoms, drugs, etc.) is a crucial task in information extraction processes especially in languages other than English. In this work, we develop a deep learning-based NLP pipeline for biomedical entity extraction in Spanish clinical narrative. Methods: We explore the use of contextualized word embeddings 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 derived from open access Spanish-language medical publications. Results: NER system integrates in-domain pre-trained Flair and FastText word embeddings, byte-pairwise encoded and the bi-LSTM-based character word embeddings. The system yielded the best performance measure with F-score of 90.84%. Error analysis showed that the main source of errors for the best model is the newly detected false positive entities with the half of that amount of errors belonged to longer than the actual ones detected entities. Conclusions: Our study shows that our deep-learning-based system with domain-specific contextualized embeddings coupled with stacking of complementary embeddings yields superior performance over the system with integrated standard and general-domain word embeddings. With this system, we achieve performance competitive with the state-of-the-art.
Corresponding Author:	Liliya Akhtyamova, MCs Technological University Dublin IRELAND
Corresponding Author E-Mail:	akhtyamova@phystech.edu
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Technological University Dublin
Corresponding Author's Secondary Institution:	
First Author:	Liliya Akhtyamova, MSc
First Author Secondary Information:	
Order of Authors:	Liliya Akhtyamova, MSc Paloma Martínez, PhD Karin Verspoor, PhD John Cardiff, PhD

Order of Authors Secondary Information:	
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Has this manuscript been submitted before to this journal or another journal in the <a href="https://www.biomedcentral.com/p/the-bmc-series-journals#journallist" target="_blank" >BMC series</ a>?	No

TECHNICAL ADVANCE ARTICLE

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

Liliya Akhtyamova[1*], Paloma Martínez[2], Karin Verspoor[3] and John Cardiff[1]

Abstract

Background: In the Big Data era there is an increasing need to fully exploit and analyse the huge quantity of information available about health. Natural Language Processing (NLP) technologies can contribute to extract relevant information from unstructured data contained in Electronic Health Records (EHR) such as clinical notes, patient's discharge summaries and radiology reports among others. Extracted information could help in health-related decision making processes. Named entity recognition (NER) devoted to detect important concepts in texts (diseases, symptoms, drugs, etc.) is a crucial task in information extraction processes especially in languages other than English. In this work, we develop a deep learning-based NLP pipeline for biomedical entity extraction in Spanish clinical narrative.

Methods: We explore the use of contextualized word embeddings 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 derived from open access Spanish-language medical publications.

Results: NER system integrates in-domain pre-trained Flair and FastText word embeddings, byte-pairwise encoded and the bi-LSTM-based character word embeddings. The system yielded the best performance measure with F-score of 90.84%. Error analysis showed that the main source of errors for the best model is the newly detected false positive entities with the half of that amount of errors belonged to longer than the actual ones detected entities.

Conclusions: Our study shows that our deep-learning-based system with domain-specific contextualized embeddings coupled with stacking of complementary embeddings yields superior performance over the system with integrated standard and general-domain word embeddings. With this system, we achieve performance competitive with the state-of-the-art.

Keywords: Natural language processing; Named entity recognition; Clinical case narratives; Deep learning; Language representations; Contextualized word embeddings; Spanish language

Background

Currently, most research in Natural Language Processing (NLP) is focused around English language texts, while textual data written in different languages are 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 native 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 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 im-

*Correspondence: akhtyamova@phystech.edu
[1]Technological University Dublin, Dublin, Ireland
Full list of author information is available at the end of the article

proved 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].

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].

Language representation models could be further applied with or without fine-tuning to a different domain problem[1]. The approach of learning on one dataset and applying the model to another dataset is called the *Transfer Learning*.

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 [17, 11].

In addition to pre-trained in-domain Spanish `FastText` embeddings [18], we generate in-domain Spanish contextualized embeddings by pre-training language representation models using the corpus retrieved from the Scientific Electronic Library Online (SciELO) website. The clinical case narrative data from the publications there was used to construct the PharmaCoNER dataset.

Our contributions 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 combining them with other types of embeddings, and compare with the standard embeddings, producing new baselines; and (4) we conduct an extensive error analysis checking the source of errors for different models.

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

Biomedical entity extraction

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

For example, Eftimov et al. [19] 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 [20]. 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 used 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 [21], rule-based techniques were actively utilized in ML and DL methods to identify patients' email addresses, locations, phone numbers, etc. In addition, participants of the challenge used domain- and language-specific gazetteers and Brown clusters derived through unsupervised ML. For example, Perez et al. [22] concluded that Brown clusters and gazetteers played a significant role in ML system performance. Further, Lopez at al. [23] tested both ML and rule-based approaches and concluded that a hybrid of the two gives the best result.

Lee et al. [24] solve the problem of biomedical NER in two steps, first discovering entities' boundaries using Support Vector Machines (SVM) techniques and then further applying ontology-based hierarchical classification method to classify identified entities. Their system got promising results 66.7% F-score on GENIA corpus [25].

Early work on machine learning-based NER includes such techniques as reranking relying on kernels [26] as well as pure feature processing [27]. Kernel-based methods for entity extraction such as SVM represented in numerous papers [28, 29, 30] overall became popular methods for extracting entities from texts including biomedical texts [31]. 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 pre-trained contextualized `Flair` embeddings fed into the 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. [32] integrated KB embeddings in their tree-structured LSTM framework, achieving approximately 3% gain in F-score.

Related to this, contextualized word embeddings together with part-of-speech (PoS) tags were examined for Bulgarian NER [33] showing sizeable improvements over state-of-the-art. In another work, a combination of different types of contextualized embeddings was explored over English biomedical literature corpora [34]. In their work, 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 [35]. They 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. [36], 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.

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. [37] placed the first with an overall F-score score of 91.05%. They used the multi-lingual large version of the pre-trained `BERT` model[2] 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 did not detail how the extracted abbreviations were further used.

The second-best results of Stoeckel et al. [38] were updated after the formal challenge with F-score 90.52%. They used the `Flair` model and made use of additional SciELO corpus, 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[3].

Sun et al. [39] placed the third-best result with F-score 89.24%. They also used the pre-trained version of `BERT` as Xiong et al. [37] with subsequent fine-tuning but without incorporating any additional features.

Overall, many participants experimented with document encoding techniques. For example, Rivera Zavala et al. [40] gathered similar size Spanish biomedical corpora to train their own `FastText` embeddings. Moreover, they used `sense2vec` [41] pre-trained embeddings. Both of these embeddings have proven useful in extracting biomedical concepts.

Methods

`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` [42], `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

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

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

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)

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 [43]. 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.

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 could 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` allowed better define 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 [44]. 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.

Additional embeddings

It has been demonstrated that the concatenation of contextualized embeddings with the standard embeddings mostly leads to the improvement of results [10, 17]. Following this, for our experiments we used the concatenation of `Flair` embeddings with Spanish general (not domain specific) `FastText` embeddings [8], in-domain Spanish biomedical `FastText` embeddings [18], byte-pairwise encoded embeddings (`BPE`) [45] and character embeddings [46]. 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 in-domain biomedical embeddings utilizing the architecture of `FastText` were trained over the SciELO[4] corpus with 100M tokens and health section of Wikipedia with 82M tokens.

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

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

Entity extraction

For PharmaCoNER, 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[5] which goes on top of Theano providing convenient way to experiment with different combinations of word embeddings. It provides off-the-shelf neural-based system supporting the 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

[4] SciELO.org

[5] https://github.com/zalandoresearch/flair

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%

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[6]. However, currently TPU does not support making predictions for downstream tasks, it is required to switch over to CPU instances for those steps.

We used a Conditional Random Fields loss [47] as it has 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 $5e^{-5}$. 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.

PharmaCoNER corpus

PharmaCoNER corpus was used for training and testing our models. It consists of 1000 annotated SPACCC articles derived from open access Spanish medical publications SciELO – an electronic library where complete full-text articles from scientific journals of Latin America, South Africa, and Spain are systematically collected and stored[7].

Table 1 shows summary statistics on 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 (11970 sentences for the merged corpora) and selected by random shuffling 10% of it for validation purposes.

Language model training set

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 BERT language representation model with the vocabulary size set to 128000.

In-domain Flair embeddings are compared with the general Flair embeddings that are a part of the Flair API. They are trained on a dump of Spanish Wikipedia dating August 2018[8].

LM training

The SciELO Flair 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 Scielo LM was done using 1 GPU instance.

The pretrained on SCiELO corpus Flair LM word embeddings we compare to that trained using a dump of Wikipedia articles with the hidden size of 2048 [9].

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 instances. For storing and uploading weights for training Google Cloud persistent storage is required. Moreover, every 8 hours Google Colab is shutting down its server, so it is needed to be resumed manually. Overall, it took more than 4 days to train BERT language representation, substantially longer than it takes to train Flair LM.

Results

The comparative results of experiments are presented in Table 2. 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.

[6] https://cloud.google.com/ml-engine/docs/tensorflow/using-tpus
[7] http://www.scielo.org
[8] https://dumps.wikimedia.org/eswiki/20180801/
[9] https://github.com/zalandoresearch/flair/issues/80

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%**

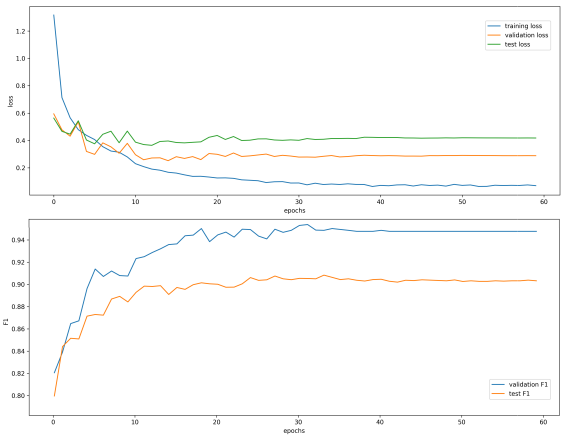

Figure 1 Training and validation loss and F-score Dependency between the number of epochs and either loss (top figure) or F1 score (bottom figure) for `Flair_Sc_ext2` model.

For the best model precision for all types of entities is higher than recall, especially for *Normalizables* entities. It means that while the model is good in determining the correct cases, it is less able to identify positive examples.

Indeed, comparing to the best systems' results it could be observed that we are superior in terms of higher precision but relatively weaker in terms of recall. Overall, our results are 0.21% behind the best system of Xiong et al. [37] 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.

Discussion

Number of training epochs Figure 1 shows an evolution of the loss and F-score over number of epochs. It could be seen that the loss becomes steady after around 27 epochs and the test F-score stabilizes at around the same time. 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.

Ablation analysis For our ablation analysis, we explored next combinations of word embeddings:

- `Flair_G_ext`: the model is trained using Spanish general domain `Flair` embeddings, 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;
- `Flair_Sc_ext2`: the extended model is trained using the custom SciELO `Flair` embeddings, SciELO `FastText` embeddings, `BPE` embeddings and character Embeddings;
- `BERT_Sc`: `BERT`-based word embeddings are trained on SciELO corpus. Then, `BERT` model is fine-tuned for the downstream task.

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 more enriched with different types of word embeddings model 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 into SNOMED-CT using Meaning Cloud tool[10], however it did not work well, as many concepts for the shared task were annotated based on their synonyms.

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

It can be seen that the majority of errors 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

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

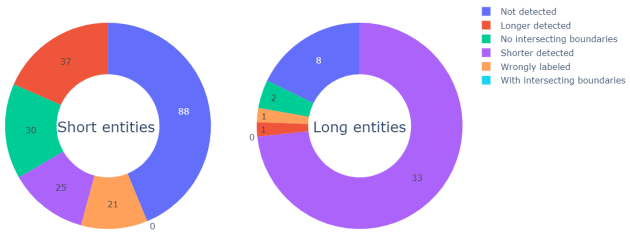

Figure 2 Distribution of errors Distribution of errors for short (less than 3 terms) and long (with length more or equal 3 terms) entities

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

are acronyms and abbreviations which could be easily missclassified 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), and 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`, 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 bigger 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 in detecting the boundaries of short biomedical concepts, often predicting longer concepts: 5 more incorrectly predicted concepts for `Flair_Sc` model and 21 more incorrectly predicted concepts for `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, `Flair_Sc` model is performing 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, there are 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 could be observed that `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.

Conclusion

In this work, we explore 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

Table 5 Examples of errors in recognizing biomedical entities by different models

		Example	Types of errors
Example 1	True 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	True 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

narratives. By leveraging the knowledge from a huge amount of unlabeled data, with language model pretraining it becomes possible to build a high-quality NER system even with this small amount of annotated data.

With this aim, we train the domain-specific Spanish language models, in particular, `Flair` and `BERT` to derive contextualized word embeddings and apply them to PharmaCoNER biomedical NER task achieving competitive results. We show that domain-specific word embeddings outperform general embeddings, despite being trained on a smaller corpus. Moreover, we show 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 could be easily mixed up with other abbreviated short words. Testing the approach analyzing other Spanish health-related texts, such as social media [48], with similar characteristics (great amount of abbreviations, lack of grammatical structure, punctuation marks, etc.) and other different ones (patient oriented terminology not included at 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 a future work.

List of abbreviations

SPACCC: Spanish Clinical Case Corpus; NLP: Natural Language Processing, NER: named entity recognition; NN: neural network; ML: machine learning; BoW: bag-of-words; SciELO: Scientific Electronic Library Online; DL: deep learning; SVM: Support Vector Machines, CNN: convolutional neural networks; RNN: recurrent neural networks; LSTM: long short-term memory networks; POS: part-of-speech; OOV: out-of-vocabulary; LM: language model; BPE: byte-pairwise encoding.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The pretrained weights for `Flair` and `BERT` models, as well as the SciELO corpora used for their training are publicly available in the Google Drive repository, http://tiny.cc/cziuiz. The dataset analysed during the current study is available on the official website of the challenge, http://temu.bsc.es/pharmaconer/.

Competing interests

The authors declare that they have no competing interests.

Funding

Part of this work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R). KV's contributions were supported by the University of Melbourne, through a Study Leave grant. LA's contributions were supported by the Technological University Dublin as part of traineeship in UC3M Spain, through Erasmus+ grant. The funders had no role in any aspect of the research, including the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Authors' contributions

LA acquired the data, designed the experiment, implemented the programming tasks, performed the analysis and drafted the paper. PMF shared ideas on solving the problem, provided feedback on the structure of paper, and revised it. KV shared ideas on solving the problem, provided the expertise in NLP and thorough revision of the paper. JC contributed to discussion of applied methods and project administration work.

Acknowledgements

Not applicable.

Author details

[1]Technological University Dublin, Dublin, Ireland. [2]Carlos III University of Madrid, Madrid, Spain. [3]The University of Melbourne, Melbourne, Australia.

References

1. Gonzalez-Agirre A, Marimon M, Intxaurrondo A, Rabal O, Villegas M, Krallinger M. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks; 2019. p. 1–10. Available from: https://github.com/PlanTL-SANIDAD/SPACCC.
2. Segura-Bedmar I, Revert R, Martínez P. Detecting drugs and adverse events from Spanish health social media streams. Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) at EACL 2014. 2014;p. 106–115.
3. Akhtyamova L, Ignatov A, Cardiff J. A Large-Scale CNN Ensemble for Medication Safety Analysis. Natural Language Processing and Information Systems NLDB 2017 Lecture Notes in Computer Science. 2017;10260.
4. Kim D, Lee J, So CH, Jeon H, Jeong M, Choi Y, et al. A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. IEEE Access. 2019;7:73729–73740. Available from: https://ieeexplore.ieee.org/document/8730332/.
5. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019. p. 2567–2577. Available from: http://arxiv.org/abs/1909.06146.
6. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. In: arXiv preprint arXiv:1301.3781; 2013. Available from: http://arxiv.org/abs/1301.3781.
7. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014;Available from: https://nlp.stanford.edu/pubs/glove.pdf.
8. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics. 2017 7;5(2307-387X):135–146. Available from: http://arxiv.org/abs/1607.04606.
9. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017 7;33(14):i37–i48. Available from: https://academic.oup.com/bioinformatics/article/33/14/i37/3953940.
10. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv preprint arXiv:180205365. 2018 2;Available from: http://arxiv.org/abs/1802.05365.
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: arXiv preprint arXiv:1810.04805; 2018. Available from: http://arxiv.org/abs/1810.04805.
12. Han X, Eisenstein J. Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019. p. 4237–4247. Available from: https://github.com/xhan77/.
13. Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018. p. 328–339. Available from: http://nlp.fast.ai/ulmfit.
14. Dai AM, Le QV. Semi-supervised Sequence Learning. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. p. 3079–3087. Available from: http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning.pdf.
15. Radford A. Improving Language Understanding by Generative Pre-Training. In: URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.; 2018. Available from: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. In Advances in neural information processing systems. 2017 6;p. 5998–6008. Available from: http://arxiv.org/abs/1706.03762.
17. Akbik A, Blythe D, Vollgraf R. Contextual String Embeddings for Sequence Labeling. In: COLING; 2018. Available from: https://github.com/zalandoresearch/flair.
18. Soares F, Villegas M, Gonzalez-Agirre A, Krallinger M, Armengol-Estapé J. Medical Word Embeddings for Spanish: Development and Evaluation. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019. p. 124–133. Available from: http://doi.org/10.5281/zenodo.2542722.
19. Eftimov T, Korouš Seljak B, Korošec P. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PLoS ONE. 2017;12(6). Available from: https://doi.org/10.1371/journal.pone.0179488.
20. Funk C, Baumgartner W, Garcia B, Roeder C, Bada M, Cohen KB, et al. Large-scale biomedical concept recognition: An evaluation of current automatic annotators and their parameters. BMC Bioinformatics. 2014 2;15(1).
21. Marimon M, Gonzalez-Agirre A, Intxaurrondo A, Rodríguez H, Antonio Lopez Martin J, Villegas M, et al. Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019); 2019. p. 618–638. Available from: https://github.com/PlanTL-SANIDAD.
22. Perez N, García-Sardiña L, Serras M, Pozo AD. Vicomtech at MEDDOCAN: Medical Document Anonymization. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019); 2019. p. 698–703. Available from: https://github.com/PlanTL-SANIDAD/SPACCC.
23. López P, D MC, Alfonso Ureña-López L, Teresa Mart M. Anonymization of Clinical Reports in Spanish: a Hybrid Method Based on Machine Learning and Rules. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019); 2019. p. 688–695. Available from: https://catalog.ldc.upenn.edu/LDC2018T01.
24. Lee KJ, Hwang YS, Kim S, Rim HC. Biomedical named entity recognition using two-phase model based on SVMs. Journal of Biomedical Informatics. 2004 12;37(6):436–447.
25. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus–a semantically annotated corpus for bio-textmining. Bioinformatics. 2003 7;19(Suppl 1):i180–i182. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btg1023.
26. Nguyen TVT, Moschitti A, Riccardi G. Kernel-based reranking for named-entity extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics (ACL); 2010. p. 901–909. Available from: https://dl.acm.org/citation.cfm?id=1944670.
27. Collins M. Ranking algorithms for named-entity extraction. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics (ACL); 2001. p. 489–496.
28. Björne J, Salakoski T. Generalizing Biomedical Event Extraction. In: Proceedings of BioNLP Shared Task 2011 Workshop; 2011. p. 183–191. Available from: http://svmlight.joachims.org/svm_.
29. Takeuchi K, Collier N. Bio-medical entity extraction using support vector machines. Artificial Intelligence in Medicine. 2005;33(2):125–137.
30. Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition. Association for Computational Linguistics (ACL); 2002. p. 1–7.
31. Patra R, Saha SK. A kernel-based approach for biomedical named entity recognition. The Scientific World Journal. 2013;2013.
32. Li D, Huang L, Ji H, Han J. Biomedical Event Extraction Based on Knowledge-driven Tree-LSTM. In: Proceedings of NAACL-HLT 2019. Association for Computational Linguistics; 2019. p. 1421–1430.
33. Simeonova L, Simov K, Osenova P, Nakov P. A Morpho-Syntactically Informed LSTM-CRF Model for Named Entity Recognition. In: arXiv preprint arXiv:1908.10261; 2019. p. preprint. Available from: http://github.com/lilia-simeonova/.
34. Sharma S, Daniel R. BioFLAIR: Pretrained Pooled Contextualized

Embeddings for Biomedical Sequence Labeling Tasks. arXiv preprint arXiv:190805760. 2019 8;Available from: http://arxiv.org/abs/1908.05760.
35. Miller T, Geva A, Dligach D. Extracting Adverse Drug Event Information with Minimal Engineering. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019. p. 22–27. Available from: https://www.aclweb.org/anthology/W19-1903.
36. Basaldella M, Collier N. BioReddit: Word Embeddings for User-Generated Biomedical NLP. In: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Association for Computational Linguistics (ACL); 2019. p. 34–38.
37. Xiong Y, Shen Y, Huang Y, Chen S, Tang B, Wang X, et al. A Deep Learning-Based System for PharmaCoNER. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks; 2019. p. 33–37. Available from: https://github.com/PlanTL-SANIDAD/SPACCC_POS-.
38. Stoeckel M, Hemati W, Mehler A. When Specialization Helps: Using Pooled Contextualized Embeddings to Detect Chemical and Biomedical Entities in Spanish. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks; 2019. p. 11–15. Available from: www.github.com/zalandoresearch/flair.
39. Sun C, Yang Z. Transfer Learning in Biomedical Named Entity Recognition: An Evaluation of BERT in the PharmaCoNER task. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks; 2019. p. 100–104.
40. Rivera Zavala RM, Martínez P. Deep neural model with enhanced embeddings for pharmaceutical and chemical entities recognition in Spanish clinical text. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks; 2019. p. 38–46. Available from: https://ufal.mff.
41. Trask A, Michalak P, Liu J. sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings. arXiv preprint arXiv:151106388. 2015 11;Available from: http://arxiv.org/abs/1511.06388.
42. Mikolov T, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. Advances in neural information processing systems. 2013;p. 3111–3119. Available from: https://arxiv.org/pdf/1310.4546.pdf.
43. Akbik A, Bergmann T, Vollgraf R. Pooled Contextualized Embeddings for Named Entity Recognition. In: NAACL; 2019. Available from: https://github.com/zalandoresearch/flair.
44. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019;(btz682). Available from: https://github.com/dmis-lab/biobert.
45. Heinzerling B, Strube M. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation ({LREC}-2018); 2018. p. 18–1473. Available from: https://aclweb.org/anthology/papers/L/L18/L18-1473/.
46. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. 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. Stroudsburg, PA, USA: Association for Computational Linguistics; 2016. p. 260–270. Available from: http://aclweb.org/anthology/N16-1030.
47. Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2001. p. 282–289. Available from: http://dl.acm.org/citation.cfm?id=645530.655813.
48. Martínez P, Martínez JL, Segura-Bedmar I, Moreno-Schneider J, Luna A, Revert R. Turning user generated health-related content into actionable knowledge through text analytics services. Computers in Industry. 2016 5;78:43–56.

List of Figures

Figures

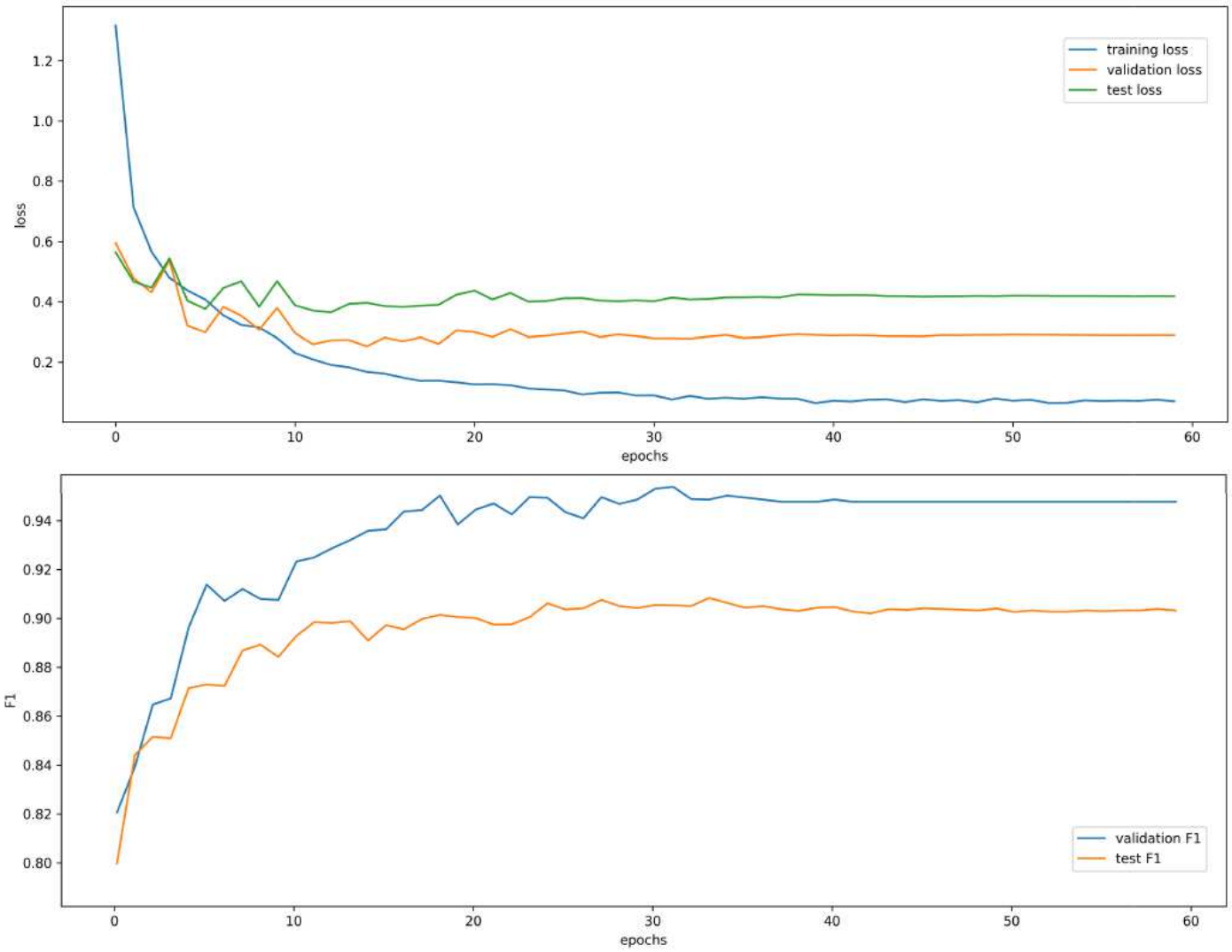

Figure 1

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.

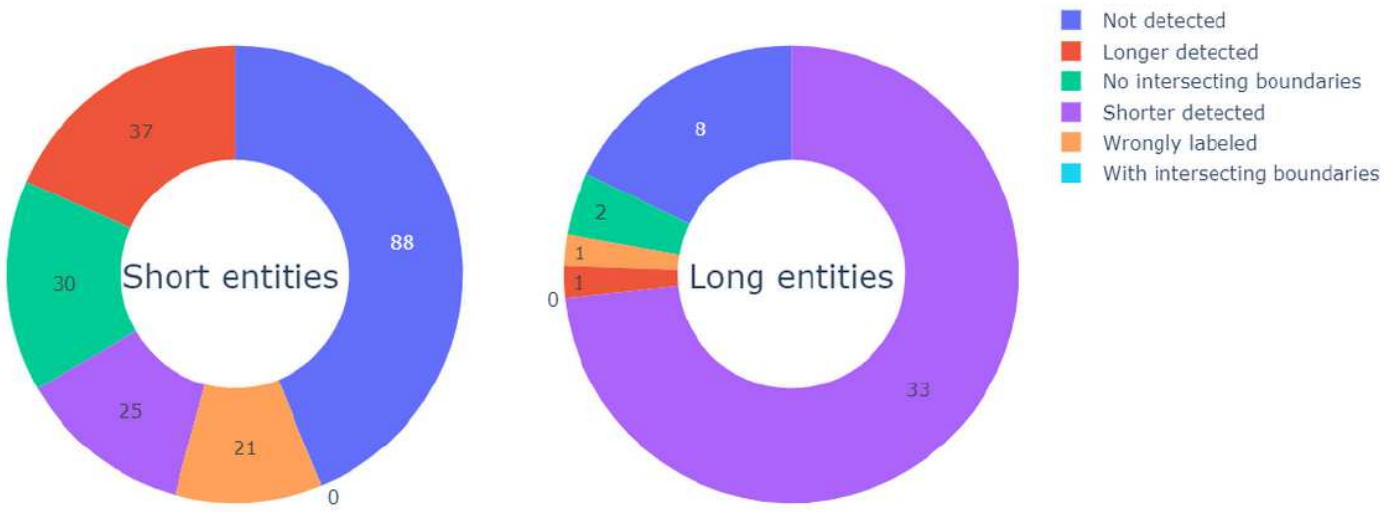

Figure 2

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