Deep neural model with enhanced embeddings for pharmaceutical and chemical entities recognition in Spanish clinical text

Renzo M. Rivera Zavala
Computer Science Department
Catholic University of Santa María
rriveraz@ucsm.edu.pe

Paloma Martínez
Computer Science Department
University Carlos III of Madrid
pmf@inf.uc3m.es

Abstract

In this work, we introduce a Deep Learning architecture for pharmaceutical and chemical Named Entity Recognition in Spanish clinical cases texts. We propose a hybrid model approach based on two Bidirectional Long Short-Term Memory (Bi-LSTM) network and Conditional Random Field (CRF) network using character, word, concept and sense embeddings to deal with the extraction of semantic, syntactic and morphological features. The approach was evaluated on the PharmaCoNER Corpus obtaining an F-measure of 85.24% for subtask 1 and 49.36% for subtask2. These results prove that deep learning methods with specific domain embedding representations can outperform the state-of-the-art approaches.

1 Introduction

Currently, the number of biomedical literature is growing at an exponential rate. Therefore, the efficient access to information on biological, chemical, and biomedical data described in scientific articles, patents, or e-health reports is a growing interest in biomedical research, industrial medicine manufacturing, and so forth. In this context, improved access to chemical and drug name mentions in biomedical texts is a crucial step downstream tasks such as drug and protein interactions, chemical compounds, adverse drug reactions, among others.

Named Entity Recognition (NER) is one of the fundamental tasks of biomedical text mining, intending to automatically extract and identify mentions of entities of interest in running text, typically through their mention offsets or by classifying individual tokens whether they belong to entity mentions or not. There are different approaches to address the NER task. Dictionary-based methods, which are limited by the size of the dictionary, spelling errors, the use of synonyms, and the constant growth of vocabulary. Rule-based methods and Machine Learning methods usually require both syntactic and semantic features as well as specific language and domain features. One of the most effective methods is Conditional Random Fields (CRF) (Lafferty et al., 2001) since CRF is one of the most reliable sequence labeling methods. Recently, deep learning-based methods have also demonstrated state-of-the-art performance for English (Hemati and Mehler, 2019; Pérez-Pérez et al., 2017; Suárez-Paniagua et al., 2019) texts by automatically learning relevant patterns from corpora, which allows language and domain independence. However, until now, to the best of our knowledge, there is only one work that addresses the generation of Spanish biomedical word embeddings (Armengol-Estapé Jordi, 2019; Soares et al., 2019).

In this paper, we propose a hybrid model combining two Bi-LSTM layers with a CRF layer. To do this, we adapt the NeuroNER model proposed in (Dernoncourt et al., 2017) for track 1 (NER offset and entity classification) of the PharmaCoNER task (Gonzalez-Agirre et al., 2019). Specifically, we have extended NeuroNER by adding context information, Part-of-Speech (PoS) tags, and information about overlapping or nested entities. Moreover, in this work, we use existing pre-trained as well as our trained word embedding models: i) a word2vec/FastText Spanish Billion Word Embeddings models (Cardellino, 2016), which were trained on the 2014 dump of Wikipedia ii) our medical word embeddings for Spanish trained using the FastText model and iii) a sense-disambiguation embedding model (Trask et al., 2015). For track 2 (concept indexing) based on the output of the previous step, we use full-text search and fuzzy matching on the SNOMED-CT Spanish Edition dictionary to obtain the corre-

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 38–46

sponding index.

Experiment results on PharmarCoNER tasks showed that our features representation improved each of separate representations, implying that LSTM-based compositions play different roles in capturing token-level features for NER tasks, thus making improvements in their combination. Moreover, the use of specific domain word vector representations (word embeddings) outperform general domain word vector and concept vector representations (concept embeddings).

2 Materials and Methods

In this section, we first describe the corpora, the training procedure and the word, concept, and sense embedding models used in our study. Then, we describe our system architecture for offset and entity classification.

2.1 Corpora

The corpus was gathered from Spanish biomedical texts from different multilingual biomedical sources:

1. The Spanish Bibliographical Index in Health Sciences (IBECS - http://ibecs.isciii.es) corpus that collects scientific journals covering multiple fields in health sciences,

2. Scientific Electronic Library Online (SciELO - https://scielo.org/es/) corpus gathers electronic publications of complete full-text articles from scientific journals of Latin America, South Africa, and Spain,

3. MedlineNLM corpus obtained from the PubMed free search engine (https://www.ncbi.nlm.nih.gov/pubmed/),

4. The MedlinePlus corpus (an online information service provided by the U.S. National Library of Medicine - https://medlineplus.gov/), consists of Health topics, Drugs and supplements, Medical Encyclopedia and Laboratory test information, and

5. The UFAL corpus (https://ufal.mff.cuni.cz/ufal_medical_corpus) is a collection of parallel corpora of medical and general domain texts.

Source corpus details are described in Table 1.

All the corpora are in XML (Dublin core format) and TXT format files. XML files were processed for extract only raw text from specific XML tags such as "title" and "description" from Spanish labels, based on the Dublin Core format as shown in Figure 1. TXT files were not processed. Raw texts from all files were compiled in a single TXT file. Texts were processed, setting all to lower, removing punctuations, trailing spaces and stop words and used as input to generate our word embeddings. Sentences pre-processing (split and tokenized) were made using Spacy [1], an open-source python library for advanced multi-language natural language processing.

2.2 Transfer Learning

Transfer learning aims to perform a task on a dataset using knowledge learned from a previous dataset (Giorgi and Bader, 2018). As shown in many works, such as speech recognition (Wang and Zheng, 2015), sentence classification (Mou et al., 2016) and Named Entity Recognition (Giorgi and Bader, 2018), transfer learning improves generalization of the model, reduces training times on the target dataset, and reduces the amount of labeled data needed to obtain high performance. In this work we used an existing generic word embedding (Word2Vec embedding trained on Spanish Wikipedia), a trained medical embedding model, and a medical/generic sense-disambiguation embedding.

Word embedding is an approach to represent words as vectors of real numbers. Word embedding models have gained much popularity among the NLP community because they are able to capture syntactic and semantic information among words. In this work, we used the Spanish Billion Words Corpora (SBWC) (Cardellino, 2016) (W2V-SBWC), which is a pre-trained model of word embeddings trained on different general domain text corpora written in Spanish (such Ancora Corpus (Martí et al., 2007) and Wikipedia) using the word2vec (Mikolov et al., 2013) implementation. The FastText-SBWC pre-trained word embeddings model was trained on the SBWC using the FastText implementation.

Furthermore, we used the sense2vec (Trask et al., 2015) model, which provides multiple dense vector representations for each word based on the

[1] https://spacy.io/

Collection\Corpus	IBECS	SciELO	MedlineNLM	MedlinePlus	UFAL
Documents	168,198	161,710	330,928	1,063	265,410
Words	23,648,768	26,169,655	4,710,191	217,515	41,604,517
Unique Words	184,936	159,997	20,942	5,099	198,424

Table 1: Biomedical Spanish corpus details.

```
<dc:description xml:lang="en">BACKGROUND Acinetobacter baumannii is an important nosocomial pathogen whose virulence
<dc:type>English Abstract</dc:type>
<dc:language>es</dc:language>
<dc:date>1998 Oct </dc:date>
<dc:title xml:lang="es">Adherencia de Acinetobacter baumannii al tejido de tráquea de la rata.</dc:title>
<dc:title xml:lang="en">[Adherence of Acinetobacter baumannii to rat tracheal tissue].</dc:title>
<dc:publisher>Revista medica de Chile</dc:publisher>
</metadata>
</record>
</pubmed-document>
```

Figure 1: Dublin core format for biomedical corpus.

sense of the word. This model is able to analyze the context of a word based on the lexical and grammatical properties of words and then assigns its more adequate vector. We used the Reddit Vector, a pre-trained model of sense-disambiguation representation vectors presented by (Trask et al., 2015). This model was trained on a collection of general domain comments published on Reddit (corresponding to the year 2015) written in Spanish and English.

2.3 Medical word and concept embeddings

We used the FastText (Bojanowski et al., 2016) implementation to train our word embeddings using the Spanish Biomedical Corpora (SBC) described in section 2.1 (FastText-SBC). Moreover, we trained a concept embedding model replacing biomedical concepts in the SBC with their unique SNOMED-CT Spanish Edition identifier (SNOMED-SBC). We used the PyMedTermino library (Lamy et al., 2015) for concept indexing. A full-text search with the Levenshtein distance algorithm (Miller et al., 2009) was applied in a first instance for concept indexing and fuzzy search with threshold using FuzzyDict implementation (Hemati and Mehler, 2019) as a second approach for concepts not found by partial matching. The FastText model uses a combination of various subcomponents to produce high-quality embeddings. It uses a standard CBOW or skip-gram models, with position-dependent weighting, phrase representations, and sub-word information in a combined manner. The training parameters for each model are shown in Table 2. Our pre-trained models can be found in Github [2] with the corpora sources, text preprocessing, and training information.

2.4 System Description

Our approach is based on a deep learning network with a preprocess step, learning transfer, two recurrent neural network layers and the last layer for CRF (see Figure 2) as proposed in (Dernoncourt et al., 2017). The input for the first Bi-LSTM layer are character embeddings. In the second layer, we concatenate character embeddings from the first layer with word, concept, and sense-disambiguate embeddings for the second Bi-LSTM layer. Finally, the last CRF layer obtains the most suitable labels for each token using a tag encoding format. For more details about NeuroNER, please refer to (Dernoncourt et al., 2017).

Our contribution consists of extending the NeuroNER system with additional features. In particular, Sense embeddigs (obtained using POS tags), concept embeddings (obtained using semantic features) and the extended BMEWO-V encoding format has been added to the network and were as a pre-preprocessing a step.

POS tags are concatenated to token in order to create dense vector representations containing word/POS information (sense embeddings) and include this in the token embedding layer of the network. Furthermore, concept features are dense vector representations generated replacing concepts with their unique SNOMED concept identi-

[2] https://github.com/rmriveraz/PharmaCoNER

Parameter\Model	FastText-SBC	SNOMED-SBC
Number of negatives sampled	20	20
Sampling threshold	6e-5	6e-5
Minimum number of word occurrences	10	10
Minimum length of character n-gram	3	3
Maximum length of character n-gram	6	6
Size of word vectors	300	300
Epochs	10	10
Processor	4 Intel Xeon 2.00 Ghz, 8 Cores, 16 Logical Processors	4 Intel Xeon 2.00 Ghz, 8 Cores, 16 Logical Processors
RAM	32 Gb	32 Gb
Corpus Size	1Gb	1Gb
Training Time	4 hours	8 hours

Table 2: Training parameters for embeddings models built in this work.

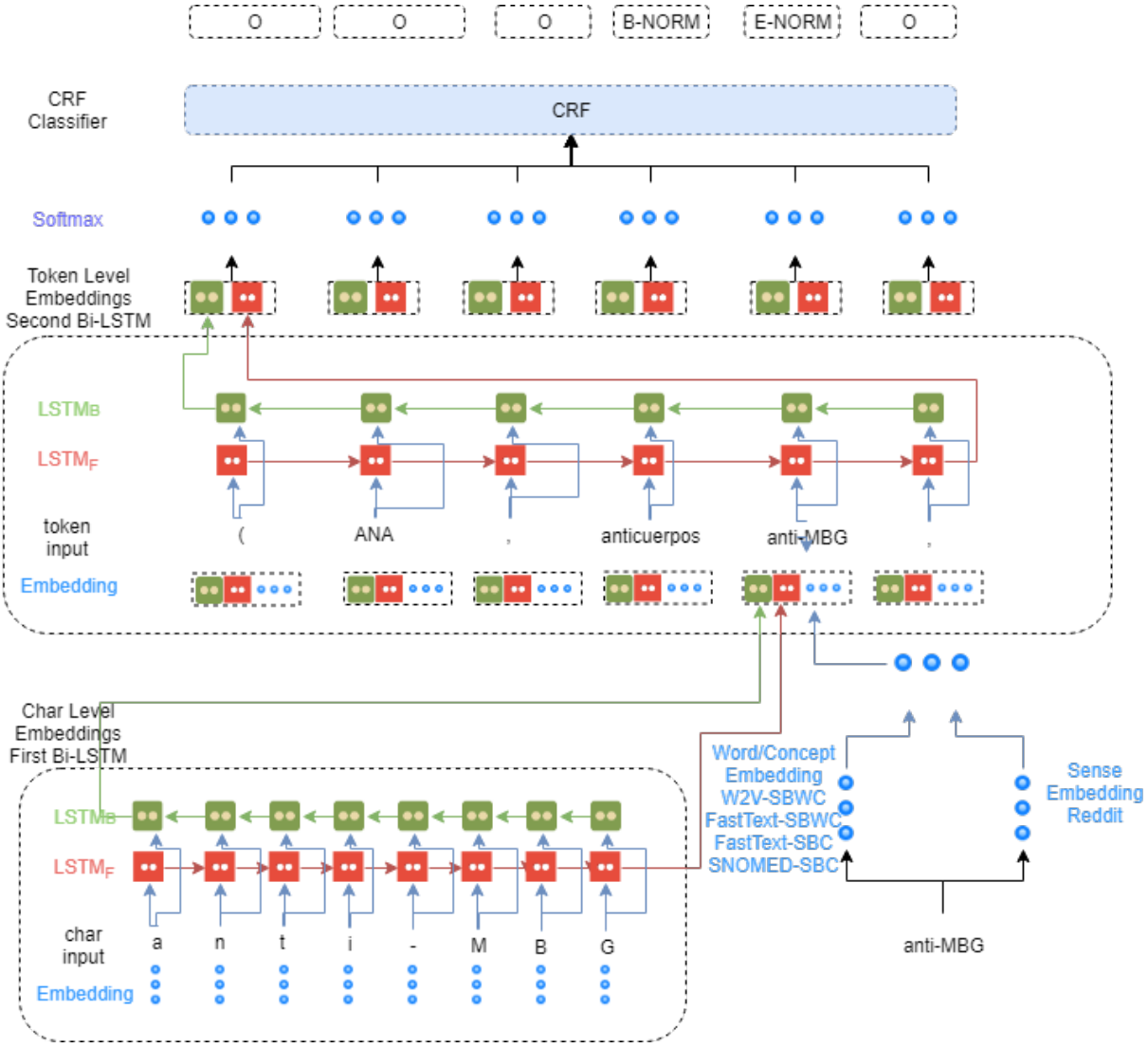

Figure 2: The architecture of the hybrid Bi-LSTM CRF model for drug and chemical compounds identifications.

fiers (concept embeddings) and include this in the token embedding layer of the network.

The BMEWO-V encoding format distinguishes the B tag for entity start, the M tag for entity continuity, the E tag for entity end, the W tag for a single entity, and the O tag for other tokens that do not belong to any entity. The V tag allows us to represent nested entities. BMEWO-V is similar to other previous encoding formarts (Borthwick et al., 1998); however, it allows the representation of nested and discontinuous entities. As a result, we obtain our sentences annotated in the CoNLL-2003 format (Tjong Kim Sang and De Meulder, 2003). An example of the BMEWO-V encoding format applied to the sentence "calcio iónico corregido 1,16 mmol/l y magnesio 1,9 mg/dl." ("ionic calcium corrected 1.16 mmol / l and magnesium 1.9 mg / dl.") can be seen in Figure 3 and Table 3.

2.4.1 First Bi-LSTM layer using character embeddings

Word embedding models are able to capture syntactic and semantic information. However, other linguistic information such as morphological information, orthographic transcription, or part-of-speech (POS) tags are not exploited. According to (Ling et al., 2015), the use of character embeddings improves learning for specific domains and is useful for morphologically rich languages. For this reason, we decided to include the character-level representations to obtain morphological and orthographic information from words. Each word is decomposed into its character n-grams and initialized with a random dense vector which is then learned. We used a 25- feature vector to represent each character. In this way, tokens in sentences are represented by their corresponding character embeddings, which are the input for our Bi-LSTM network.

2.4.2 Second Bi-LSTM layer using word and Sense embeddings

The input for the second Bi-LSTM layer is the concatenation of character embeddings from the first layer with the pre-trained word or concept embeddings and sense-disambiguation embeddings (described in sections 2.2 and 2.3) of the tokens in a given input sentence. The second layer goal is to obtain a sequence of probabilities for each tag in the BMEWO-V encoding format. In this way, for each input token, this layer returns six probabilities (one for each tag in BMEWO-V). The final tag should be with the highest probability for each token.

2.5 Last layer based on Conditional Random Fields (CRF)

To improve the accuracy of predictions, a Conditional Random Field (CRF) (Lafferty et al., 2001) model is trained, which takes as input the label probability for each independent token from the previous layer and obtains the most probable sequence of predicted labels based on the correlations between labels and their context. Handling independent labels for each word shows sequence limitations. For example, considering the drug sequence labeling problem an "I-NORMALIZABLES" tag cannot be found before a "B- NORMALIZABLES" tag or a "B- NORMALIZABLES" tag cannot be found after an "I-NORMALIZABLES" tag. Finally, once tokens have been annotated with their corresponding labels in the BMEWO-V encoding format, the entity mentions must be transformed into the BRAT format. V tags, which identify nested or overlapping entities, are generated as new annotations within the scope of other mentions.

3 Evaluation

As it was described above, our system is based on a deep network with two Bi-LSTM layers and the last layer for CRF. We evaluate our NER system using the train, validation, and test datasets (SPACCC) provided by the PharmaCoNER task organizers (Gonzalez-Agirre et al., 2019). Detailed information for each datasets can be seen in Table 4. The PharmacoNER dataset is a manually annotated corpus of 1,000 clinical cases written in Spanish and annotated with mentions of chemical compounds, drugs, genes, and proteins. The dataset consists of Normalizables (4,398), No Normalizables (50), Proteins (3,009), and Unclear (167) labels. Further details can be found in (Gonzalez-Agirre et al., 2019).

The PharmaCoNER task considers two subtasks. Track 1 consider offset recognition and entity classification of pharmacological substances, compounds, and proteins. Track 2 considers concept indexing where for each entity, the list of unique SNOMED concept identifiers must be generated. Scope level F-measure is used as the main metric where true positives are entities which match with the gold standard clue words and scope

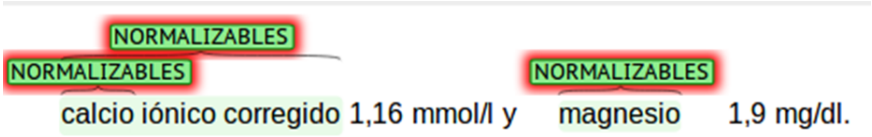

Figure 3: BRAT annotation example from PharmaCoNER corpus sentence.

token	start offset	end offset	tag	tag
calcio	0	6	V-NORMALIZABLES	W-NORMALIZABLES
iónico	8	14	V-NORMALIZABLES	O
corregido	16	25	V-NORMALIZABLES	O
1,16	27	31	O	O
mmol/l	33	39	O	O
y	41	42	O	O
magnesio	43	51	V-NORMALIZABLES	O
1,9	52	55	O	O
mg/dl	57	62	O	O
.	63	64	O	O

Table 3: Tokens annotated with BMEWO-V encoding in the ConLL-2003 format.

Dataset	Subset	Documents	Sentences	Entities
PharmaCoNER	Train	500	8036	3822
	Valid	250	3759	1926
	Test	3751	62000	

Table 4: PharmaCoNER subsets details.

boundaries assigned to the clue word. A detailed description of evaluation can be found in the PharmaCoNER web [3].

3.1 Track 1 - Offset detection and Entity Classification

The NER task is addressed as a sequence labeling task. For track 1 we tested different configurations with various pre-trained embeddings models. The embedding models and their parameters are summarized in Table 5. Table 6 describes our different experiments configurations.

In Table 8, we compare the different pre-trained models in Spanish on the validation dataset. As shown in Table 8 specific domain word embeddings outperform general domain models by almost 5 points. For the test dataset, we applied our best system configuration FastText-SBC + Reddit (see Table 8) obtaining an f-score of 85.24% for offset detection and entity classification. Furthermore, Table 7 shows the classification results obtained by our best system configuration for track 1 with a micro average of 88.10% for valid dataset.

Moreover, we compared our best system configuration (FastText-SBC + Reddit) with the baseline system (NeuroNER without POS and BMEWO-V format encoding) using the same pre-trained models and configuration. Table 9 shows that our extended system outperforms the baseline system, which has proven that POS and BMEWO-V format to be an additional source of information that can be leveraged by neural networks and keep our model domain agnostic. Furthermore, the use of specific domain word embeddings highly improve performance as shown in Table 8.

3.2 Track 2 - Concept Indexing

For track 2, we applied the same approach described for SNOMED-SBC model training in section 2.3 for entities obtained in the previous task. We used the PyMedTermino library employing a two-stage search using full-text search and fuzzy search for concepts not found by partial matching. Table 10 shows our result for valid and test dataset

[3] http://temu.bsc.es/pharmaconer/index.php/evaluations

Detail	W2V-SBWC	FastText-SBWC	FastText-SBC	SNOMED-SBC	Reddit
Type	Word	Word	Word	Concept	Sense
Corpus size	1.5 billion	1.5 billion	6 trillion	6 trillion	2 billion
Vocab size	1 million	1 million	2 million	2 million	1 million
Array size	300	300	300	300	128
Algorithm	Word2Vec Skip-gram BOW	FastText Skip-gram BOW	FastText Skip-gram BOW	FastText Skip-gram BOW	Sense2Vec

Table 5: Embedding models details.

Parameter	Run 1	Run 2	Run 3	Run 4
Sense-disambiguation embedding dimension	128	128	128	128
Pre-trained word embeddings	FastText-SBC + Reddit	W2V-SBWC + Reddit	FastText-SBWC + Reddit	SNOMED-SBC + Reddit
Word embeddings dimension	300	300	300	300
Character embedding dimension	50	50	50	50
Hidden layers dimension (for each LSTM)	100	100	100	100
Learning method	SGD	SGD	SGD	SGD
Dropout rate	0.5	0.5	0.5	0.5
Learning rate	0.005	0.005	0.005	0.005
Epochs	100	100	100	100

Table 6: System hyperparameters for each run.

Entity	Precision (%)	Recall (%)	F-score (%)
Normalizables	92.38	86.41	89.29
No_Normalizables	0.00	0.00	0.00
Proteins	93.29	85.35	89.14
Unclear	87.80	70.59	78.26
Micro-average	91.75	84.74	88.10

Table 7: Results for valid dataset entities.

Experiment	Embedding Model	Precision (%)	Recall (%)	F-score (%)
Run 4	SNOMED-SBC + Reddit	83.52	74.97	79.02
Run 2	W2V-SBWC + Reddit	83.85	75.75	79.60
Run 3	FastText-SBWC + Reddit	84.70	77.31	80.84
Run 1	FastText-SBC + Reddit	**89.13**	**82.61**	**85.75**

Table 8: Embeddings model results for track 1 on valid dataset.

for track 2.

Our results for track 2 are low due to a large number of misspellings that exceed the similarity threshold such as ”diacepam” (”diazepam”), drug names where the identifier corresponds to the active substance as ”durogesic” (”Duragesic”) active ingredient ”fentanyl” (”fentanyl”), identifiers not existing in SNOMED CT, such as CHEBI:135810

System	Precision (%)	Recall (%)	F-score (%)
NeuroNER	86.38	82.07	84.16
Extended NeuroNER	**89.13**	**82.61**	**85.75**

Table 9: Baseline comparison for track 1 on valid dataset.

and 373757009 and false positives, such as diseases identified as NORMALIZABLE entities and PROTEIN tokens not annotated in the corpus.

4 Conclusions

In this work, we propose a system for the detection of chemical compounds, drugs, genes, and proteins in clinical narrative written in Spanish. We address the named entity recognition task as a sequence labeling task. Our hybrid model based on machine and deep learning approaches only use dense vector representations features instead of hand-crafted word-based features. We proved that as in other tasks such as NER, the use of dense representation of words such as word-level embeddings, character-level embeddings, and sense embeddings are helpful for named entity recognition. The hybrid system achieves satisfactory performance with F-score over 85%. The extension of NeuroNER network is domain-independent and could be used in other fields, although generic prebuilt word embeddings are used, new medical Spanish word and concept embeddings have been generated for this work.

As future work, we plan to enhance the SNOMED-CT concept embeddings and analyze why its performance is lower than the medical word embeddings. We plan to test whether other supervised classifiers such as Markov Random Fields, Optimum-Path-Forest, or CRF as RNN would obtain more benefit from dense vector representation. That is to say, we would use the same continuous representations with the after-mentioned classifiers. Apart from that, we could train word embeddings obtained from multiple multilingual biomedical corpus to obtain multilingual word representations and test other word representation algorithms such as concept embeddings using UMLS or other biomedical unique concept identifier dictionary. The motivation would be to see whether word embeddings generated with multilingual biomedical domain texts can help to improve the results and provide a deep learning model language and domain-independent.

Funding

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).

References

Marimon Montserrat Krallinger Martin Armengol-Estapé Jordi, Soares Felipe. 2019. Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts. *Genomics Inform*, 17(2):e15–.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Sixth Workshop on Very Large Corpora*.

Cristian Cardellino. 2016. Spanish Billion Words Corpus and Embeddings.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102, Copenhagen, Denmark. Association for Computational Linguistics.

John M Giorgi and Gary D Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the BioNLP Open Shared Tasks (BioNLP-OST)*, pages 1–X, Hong Kong, China. Association for Computational Linguistics.

Wahed Hemati and Alexander Mehler. 2019. Lstmvoter: chemical named entity recognition using a conglomerate of sequence labeling tools. *Journal of Cheminformatics*, 11(1):3.

Dataset	Precision (%)	Recall (%)	F-score (%)
valid	51.72	50.57	51.14
test	50.00	49.28	49.64

Table 10: Results for PharmaCoNER track 2 on valid and test dataset.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jean Baptiste Lamy, Alain Venot, and Catherine Duclos. 2015. PyMedTermino: An open-source generic API for advanced terminology services. *Studies in Health Technology and Informatics*, 210:924–928.

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal. Association for Computational Linguistics.

M.A. Martí, M. Taule, M. Bertran, and L. Márquez. 2007. AnCora: Multilingual and Multilevel Annotated Corpora.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? *CoRR*, abs/1603.06111.

Martín Pérez-Pérez, Obdulia Rabal, Gael Pérez-Rodríguez, Miguel Vazquez, Florentino Fdez-Riverola, Julen Oyarzábal, Alfonso Valencia, Anália Lourenço, and Martin Krallinger. 2017. Evaluation of chemical and gene/protein entity recognition systems at biocreative v.5: the cemp and gpro patents tracks.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Víctor Suárez-Paniagua, Renzo M. Rivera Zavala, Isabel Segura-Bedmar, and Paloma Martínez. 2019. A two-stage deep learning approach for extracting entities and relationships from medical texts. *Journal of Biomedical Informatics*, 99:103285.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Andrew Trask, Phil Michalak, and John Liu. 2015. sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. *CoRR*, abs/1511.06388.

Dong Wang and Thomas Fang Zheng. 2015. Transfer learning for speech and language processing. *CoRR*, abs/1511.06066.