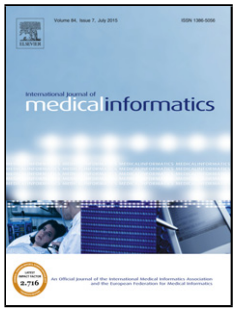


Title: Measuring the Effect of Different Types of Unsupervised Word Representations on Medical Named Entity Recognition

Author: Arantza Casillas Nerea Ezeiza Iakes Goenaga Alicia Pérez Xabier Soto

PII: S1386-5056(18)31031-1
DOI: https://doi.org/doi:10.1016/j.ijmedinf.2019.05.022
Reference: IJB 3886

To appear in: *International Journal of Medical Informatics*

Received date: 17 September 2018
Revised date: 7 March 2019
Accepted date: 21 May 2019

Please cite this article as: Arantza Casillas, Nerea Ezeiza, Iakes Goenaga, Alicia Pérez, Xabier Soto, Measuring the Effect of Different Types of Unsupervised Word Representations on Medical Named Entity Recognition, *<![CDATA[International Journal of Medical Informatics]]>* (2019), https://doi.org/10.1016/j.ijmedinf.2019.05.022

Measuring the Effect of Different Types of Unsupervised Word Representations on Medical Named Entity Recognition

Arantza Casillas[a,*], Nerea Ezeiza[a], Iakes Goenaga[a], Alicia Pérez[a], Xabier Soto[a]

[a]*IXA group, University of the Basque Country (UPV-EHU)*
Manuel Lardizabal 1, 20080 Donostia, Spain

Highlights

- Medical Entity Recognition is crucial for accurate clinical text processing
- Our approach implements neural networks and word embeddings
- The focus is on robust dense word representations
- This work serves as a guide to choose the right corpora, algorithm and parameters

*Corresponding author
Email addresses: arantza.casillas@ehu.eus (Arantza Casillas), n.ezeiza@ehu.eus (Nerea Ezeiza), iakes.goenaga@ehu.eus (Iakes Goenaga), alicia.perez@ehu.eus (Alicia Pérez), xabier.soto@ehu.eus (Xabier Soto)

Preprint submitted to International Journal of Medical Informatics *January 10, 2019*

Measuring the Effect of Different Types of Unsupervised Word Representations on Medical Named Entity Recognition

Arantza Casillas[a,*], Nerea Ezeiza[a], Iakes Goenaga[a], Alicia Pérez[a], Xabier Soto[a]

[a]*IXA group, University of the Basque Country (UPV-EHU)*
Manuel Lardizabal 1, 20080 Donostia, Spain

Abstract

Bacground: This work deals with Natural Language Processing applied to the clinical domain. Specifically, the work deals with a Medical Entity Recognition (MER) on Electronic Health Records (EHRs). Developing a MER system entailed heavy data preprocessing and feature engineering until Deep Neural Networks (DNNs) emerged. However, the quality of the word representations in terms of embedded layers is still an important issue for the inference of the DNNs.

Goal: The main goal of this work is to develop a robust MER system adapting general-purpose DNNs to cope with the high lexical variability shown in EHRs. In addition, given that EHRs tend to be scarce when there are out-domain corpora available, the aim is to assess the impact of the word representations on the performance of the MER as we move to other domains. In this line, exhaustive experimentation varying information generation methods and network parameters are crucial.

Methods: We adapted a general purpose sequential tagger based on Bi-directional Long-Short Term Memory cells and Conditional Random Fields (CRFs) in order to make it tolerant to high lexical variability and a limited

*Corresponding author
Email addresses: arantza.casillas@ehu.eus (Arantza Casillas), n.ezeiza@ehu.eus (Nerea Ezeiza), iakes.goenaga@ehu.eus (Iakes Goenaga), alicia.perez@ehu.eus (Alicia Pérez), xabier.soto@ehu.eus (Xabier Soto)

amount of corpora. To this end, we incorporated part of speech (POS) and semantic-tag embedding layers to the word representations.

Results: One of the strengths of this work is the exhaustive evaluation of dense word representations obtained varying not only the domain and genre but also the learning algorithms and their parameter settings. With the proposed method, we attained an error reduction of 1.71 (5.7%) compared to the state-of-the-art even that no preprocessing or feature engineering was used.

Conclusions: Our results indicate that dense representations built taking word order into account leverage the entity extraction system. Besides, we found that using a medical corpus (not necessarily EHRs) to infer the representations improves the performance, even if it does not correspond to the same genre.

Keywords: Electronic Health Records, Medical Named Entity Recognition, Health Information Systems, Neural Network

1. Introduction

This work deals with automatic information extraction from medical health records by means of Natural Language Processing (NLP). Information extracted from medical texts has been proven to be helpful in many clinical practices [1] such as classification of Electronic Health Records (EHRs) [2], automatically building patient trajectories [3] [4] [5] [6], or finding Adverse Drug Reactions [7]. Needless to say, Medical named Entity Recognition (MER), e.g., detecting instances of diseases or drugs, is a crucial step towards accurate medical text processing in downstream applications such as those aforementioned. Specifically, in the context of this project, one of the goals is to retrieve relations between drugs and diseases, with special interest on harmful or potentially harmful reactions, such as adverse drug reactions, but it is also essential to identify diseases to improve the classification of EHRs according to the pathology.

Extracting information from clinical records is challenging in comparison to other texts. Note that EHRs are highly noisy with an intensive usage of acronyms and abbreviations (3,774 short forms were found in 99 EHRs [8]) as

well as a high variability of terminology [9]. In fact, Leaman et al. [9] identified high term variation as the primary cause of low performance in EHRs.

Medical NER represents a core-element in downstream information extraction applications. Extracting critical information about previously prescribed treatments (i.e. drugs) and associating them with past episodes (diseases) can be useful to summarize the health condition of patients and to support decision making. For example, discovering diagnostic terms is crucial for further classification according to the International Classification of Diseases [10], and recognizing drugs and diseases for Adverse Drug Reaction extraction in pharmacosurveillance [11]. At this stage, we work on drugs and diseases, as they are key elements in medical information extraction applications and, to that end, we have already developed some resources. Needless to say, the work can be extended to other kinds of entities in the future (e.g. body parts, dosages, dates etc.).

Our main **contributions** are:

1. We developed a machine learning system for discovering medical entities from spontaneous clinical narrative[1], modeling the problem as a sequential tagging task.
2. We evaluated neural models using three corpora and five different approaches in order to learn from them in an unsupervised manner, and also compare the performance of the model with respect to a number of parameter settings.

2. Related Work

General domain Named Entity Recognition (NER) has been addressed using different techniques, such as Conditional Random Fields (CRF) [13] or Support

[1] By spontaneous we refer to the "real-time" notion expressed by the World Health Organization in [12] for the definition of electronic medical record (EMR) as a real-time patient health record.

Vector Machines (SVM) [14]. Recently, Deep Neural Networks (DNN) have been successfully applied in general domain NER [15], [16], [17]. For training, big quantities of raw data are needed to obtain dense word representations and a minimal quantity of manually annotated corpus for training. The main contribution of DNN in these tasks depends on:

- Simplifying the recognition process by not requiring preprocessing, feature engineering, or the application of linguistic analysis tools [18].
- Inducing knowledge from big amounts of raw corpora, thus breaking the knowledge acquisition bottleneck [19]. One extended idea is that bigger quantities of data result in better dense word representations (i.e., embeddings [20]) and, consequently, many researchers employ off-the-shelf pretrained word embeddings over huge quantities of unlabeled text. Using such pretrained embeddings limits the possibilities of exploring the effect different hyperparameters or corpus domain might have on the results. In this work we delve into this question empirically.

Looking at the literature, some works aim to explore the in and out domain effect. For example, Roberts [21] presented a series of experiments to evaluate how to combine small-but-representative corpora and large-but-unrepresentative corpora for building word embeddings for clinical Natural Language Processing (NLP) tasks; they concluded that combining multiple corpora is the safest option.

Lai et al. [22] concluded that it is not clear whether the domain is more important than the corpus size because the impact depends on the task: a task involving semantics such as text classification might be more influenced by the domain than more syntactically oriented tasks such as part-of-speech (POS) tagging.

In our case, we delve into this question in a domain with high variability and very little free available data, namely, EHRs. For general NER, Lample et al. [16] proposed a neural architecture based on bidirectional Long-Short Term Memory (LSTM) and CRFs using no language-specific features for the detection

of entities, with state-of-the-art results. Melamud et al. [23] studied the role of context and dimension on the effectiveness of different word embeddings for different language processing tasks, concluding that it is crucial to choose the optimal context (window) and vector dimensionality to get the best results in specific tasks.

Regarding the biomedical domain, Chiu et al. [24] studied how several parameters might affect the performance of dense word representations in Biomedical NLP, using exclusively the Word2Vec tool, although they tested these parameters independently. Indeed, they suggested that their combinations should be analyzed. They concluded that extending the vector dimensions over 200 brings no improvement on MER, although there are still improvements in some cases with higher values.

Jagannatha and Yu [25] and Jagannatha and Yu [26] used different Recurrent Neural Networks to pursue MER. They obtained their embeddings applying Word2Vec's skip-gram algorithm (SkipG henceforth) over PubMed articles, English Wikipedia and 100,000 EHRs. They reported an F-score of 0.72 over diseases and 0.90 over drugs.

In a more recent work, Xu et al. [27] used a model based on a bidirectional LSTM and Conditional Random Field for Medical NER. Their model contains three layers and relies on character-based word representations, managing to obtain a 0.80 F-score on a NCBI Disease Corpus.

Motivated by the antecedents, several issues require further investigation in order to gain an insight into the relevant aspects that might be determining:

- The **selection of the right corpus** becomes crucial, especially in specific domains such as health. One open question is to which extent big quantities of general purpose corpora help to overtake domain limitations [22], [28].

- Another question is how **different methods and parameters** for obtaining word representations may influence the results.

3. Materials and Methods

In line with recent work done on general NER [16], we developed a MER system in two phases (see Figure 1):

1. Obtain unsupervised dense word representations from large unannotated corpora.
2. Apply neural networks (Bi-LSTM + CRF) to infer the MER model.

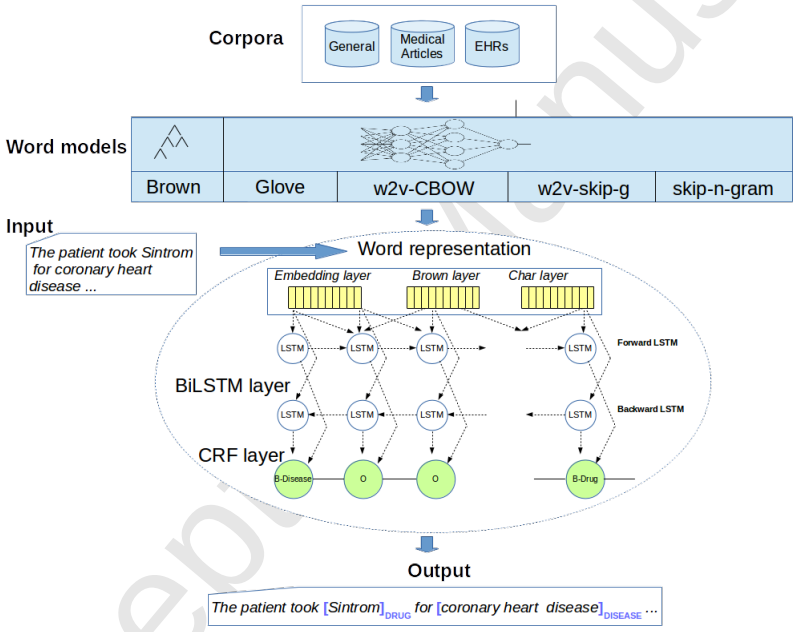


Figure 1: General architecture of the system.

3.1. Large unannotated Corpora

To measure the impact of the corpus type and domain on the quality of the **embeddings** we used three different corpora:

- *General Corpus* (GEN henceforth). This corpus consists of around 1.5 billion words, compiled from different corpora, treebanks and resources from the web (`http://crscardellino.github.io/SBWCE/`).

- *Medical Online Corpus* (GEN-MED). We compiled around 1.3 million words from several sources: Spanish documents from the Spanish-English parallel UFAL Medical Corpus 1.0, documents obtained from crawling the Spanish version of MedLine and drug information web pages, as well as a subset of Wikipedia articles filtered using SNOMED-CT concepts.

- *IXAMed Spanish EHR Corpus* (EHR). The IXAMed corpus comprises 141,800 EHRs collected over 4 years at Galdakao and Basurto Hospitals. It contains 52 million tokens.

In order to preprocess the documents, we have made use of FreeLing-Med [29], which is a variant of FreeLing [30] focused on medical texts. By enriching the lexica of the Freeling analyzer with biomedical terms extracted from dictionaries and ontologies such as SNOMED-CT, FreeLing-Med is able to automatically detect medical terms in texts.

3.2. Manually tagged Corpus

A subset of 121 EHRs were manually annotated by experts of the Basque health network with, among others, two kinds of medical entities: diseases and drugs. The annotation process was divided into three phases [31]: i) Definition of the scope of the annotation process and creation of the work team ii) Preliminary annotation iii) Annotation of the corpus with the help of FreeLing-Med. Two experts took part in the annotation and 90.53% inter-annotator agreement was achieved. As pointed out by Pradhan et al. [32], *traditional agreement measures such as Cohen's κ and Krippendorf's α are not applicable for measuring agreement for entity mention annotation*, because what both measures do is to calculate the inter-annotator agreement discounting the probability of agreement by chance. In our case, given that we focus on drugs and diseases, most of the tokens do not receive any annotation and, consequently, the probability of agreeing by chance would be close to zero. Therefore, we have directly measured the percentage of entities tagged in the same way by both annotators.

Finally, the annotated corpus was randomly divided into the usual training, development and testing folds (see Table 1).

Corpus	words	sentences	drug	disease
train	37,286	4,076	884 (925)	2,319 (2,895)
dev	16,478	1,745	522 (547)	1,043 (1,319)
test	14,458	1,429	456 (467)	934 (1,209)

Table 1: Manually tagged corpus in terms of words, sentences, number of drug and disease entity-tokens and words in parenthesis (indicating that tokens tend to be multi-word units).

3.3. Deep Learning (neural network model)

LSTM [33] neural networks are suited for sequential data labeling. Basically, they take a sequence input $(x_1, x_2, x_3, \cdots, x_n)$ obtaining the corresponding output $(h_1, h_2, h_3, \cdots, h_n)$ at each time step. Through their gate based system LSTMs are able to automatically regulate how much of the previous context should persist and how much should be renewed. Bi-LSTMs are a special case of LSTM where two LSTM nets are employed, treating the input sequence from left to right and from right to left (forward and backward LSTM). This work is based on the implementation by Lample et al. [16]. In order to enrich their approach, we have incorporated the ability to accept multiple dense word representations or embeddings, as well as additional information layers (such as Brown clusters, POS or semantic tags) at each time step. With these additional layers and word representations, our approach is able to outperform their setup in medical domain. Thus, these results suggest that the abstract structural or semantic information brought by this third layer synergized with the previously proposed two surface layers and, apparently, complemented each other leading to significant improvements.

3.4. Word Representations

Word dense representations capture the semantic and syntactic information of words from large unlabeled corpora. There are basically three methods to

obtain them: dimensionality reduction count-based methods such as Singular Value Decomposition (SVD) [34], predictive methods mostly based on neural networks (Word2Vec [20]) and clustering methods where a word is represented by the cluster it belongs to in a language model (Brown clusters [35]). We employed GloVe as a count-based method, Word2Vec in its two variants, SkipNG as prediction method and, finally, Brown clusters.

GloVe [36] (for Global Vectors) captures word meaning by means of the ratio of co-occurrence probability of the words in the corpus. Word2Vec[2] [20] attains a dense word representation by using neural nets to predict the next word given context words within a window (SkipG variant), or predicting context words within a window size given a word (CBOW variant). As a way to simplify the system, SkipG selects one word from the window of the target word using its embedding as a context representative. CBOW, conversely, simplifies obtaining the context vector by averaging over the embeddings of context words. SkipNG [37] can be seen as a variant of CBOW where the context vector is computed not by averaging over the context word embeddings but as a weighted sum of the individual word embeddings. Each word embedding weight is assigned depending on the position of each word.

On the other hand, the randomly initialized character lookup table contains an embedding for each character. The character embeddings corresponding to every character in a word are given in direct and reverse order to a forward and a backward LSTM. The final representation for a word is obtained from its character embeddings by concatenating the forward and backward representations derived from the bidirectional LSTM.

Brown clustering [35] is an agglomerative clustering algorithm that builds a class-based language model based on word co-occurrences. It is an iterative algorithm, where at each iteration two clusters are merged by choosing from all possible mergeable clusters the pair that produces the smallest decrease in the corpus likelihood.

[2]We have used the default settings for parameters not mentioned in the manuscript.

4. Results

First we measured the effect that different pretrained word representations have on MER, depending on the corpora type and algorithm employed to obtain them (see Figure 2). For example, dimension 300 and window size 5 are the hyperparameters employed to obtain the following freely available pretrained embeddings: Spanish Billion Words Corpus[3], GloVe[4] and Word2Vec[5].

One remarkable aspect is the effect of including a character LSTM that can be helpful in the NER task, as these character networks can be useful to generalize over strings such as prefixes or suffixes, especially useful for detecting entities (such as the *-itis* or *-osis* suffixes for diseases). Table 2 gives the results for each corpus type and dimensions using the best method (SkipNG), with and without including a character LSTM layer.

Embed. source	Dim. and WinSize	Char. layer	
		w/o	with
GEN	D50-W1	64.95	69.02 (+4.07)
	D300-W5	65.85	69.19 (+3.34)
GEN-MED	D50-W1	66.75	70.58 (+3.83)
	D300-W5	65.99	70.46 (+4.47)
EHR	D50-W1	71.90	73.67 (+1.77)
	D300-W5	70.99	72.86 (+1.87)

Table 2: Impact of the addition of the character layer on alternative spaces obtained from different corpora measured in terms of F-score on the SkipNG algorithm for different dimensions (Dim) and window sizes (WinSize).

The system taking as input the SkipNG pretrained embeddings from EHRs outperformed the rest. This setting was selected for the following round of

[3] crscardellino.me/SBWCE
[4] nlp.stanford.edu/projects/glove/
[5] code.google.com/archive/p/word2vec

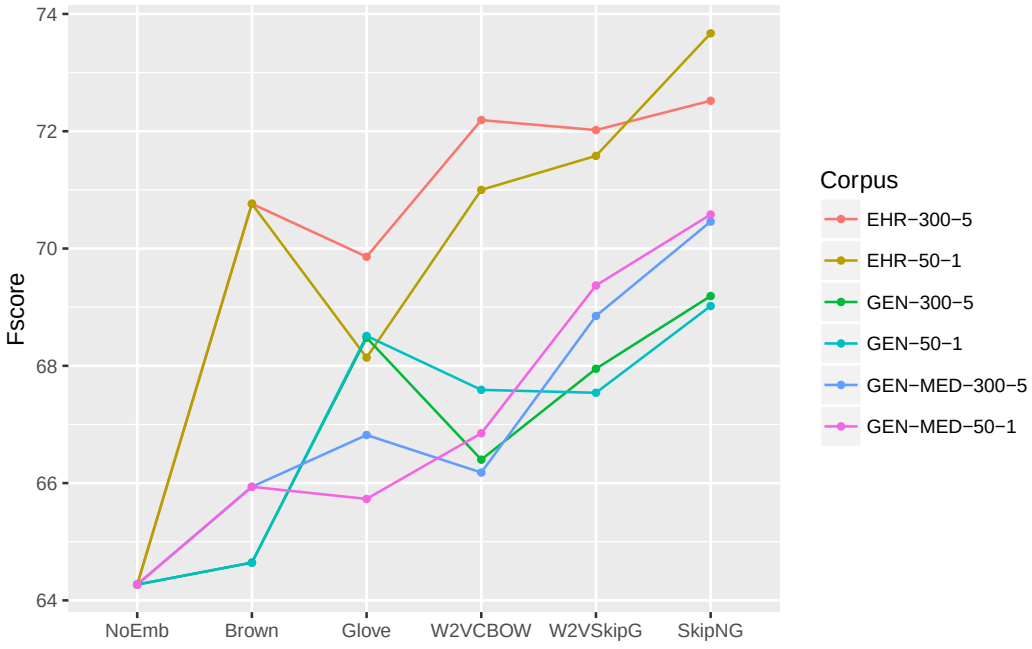


Figure 2: Medical Entity Recognition for different types of corpora, neural network dimensions and type of unsupervised knowledge employed (development set). The figure reports the results employing no pretrained embeddings, and pretrained word representations learned from generic corpora (GEN), generic medical corpora (GEN-MED) and EHRs using Brown clusters, GloVe, CBOW, SkipG and SkipNG. We set the embedding dimension to 300 with window size 5 (denoted as 300-5), and embedding dimension of 50 with window size of 1 (denoted as 50-1). The baseline (NoEmb) corresponds to the system without pretrained embeddings, that is, calculating the embeddings based only on the reduced vocabulary present in the annotated training set.

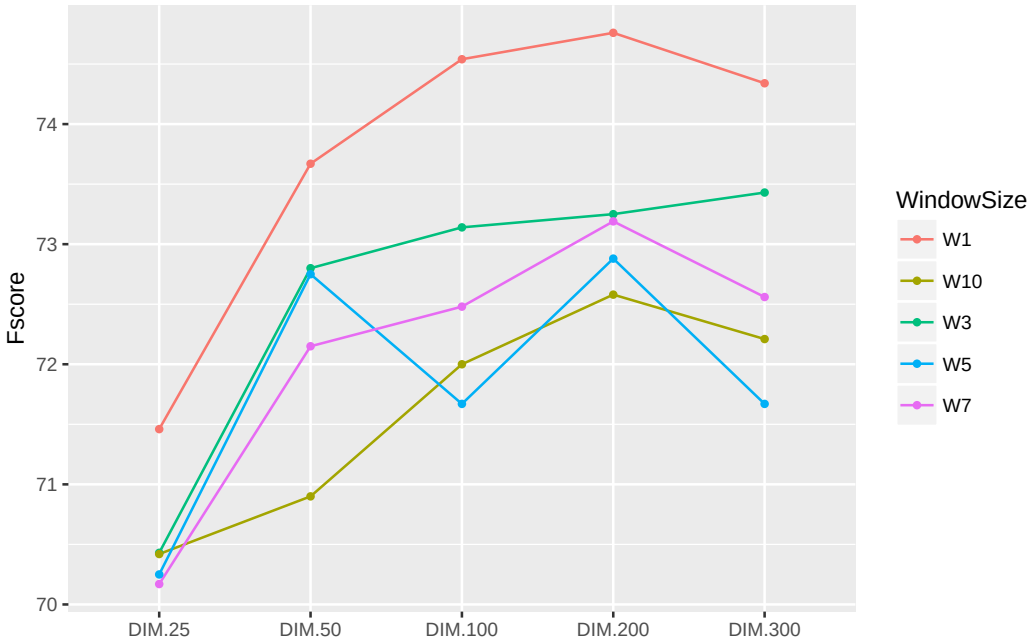


Figure 3: Effect of window size and dimension of embeddings (F-score for EHR embeddings and SkipNG).

experiments, where the impact of varying the dimension of the word embeddings (25, 50, 100, 200 and 300) and window size (taking a window of 1, 3, 5, 7, and 10) was measured. Figure 3 illustrates the effect of jointly varying the embedding dimension and window size on the best performing algorithm (SkipNG) trained with EHRs, as proposed by Chiu et al. [24]. We can see that the smallest window of size 1 outperforms the other types. With respect to higher context window values, the performance of the system varies depending on the size of the word embedding dimension. Our impression is that in this specific domain lengthy contexts are not frequently repeated and, hence, the training does not lead to reliable results. Besides, the lexical variability in this domain requires of models with high generalization ability, as provided by W1. Regarding the dimension of the embeddings, the best result corresponds to a dimension of 200.

Finally, Table 3 gives the results on the test set, presenting them together with the current state-of-the-art using standard machine learning algorithms (SVM, Perceptron and CRF [38]). The first line presents the baseline when using only wordforms, together with feature engineering (such as capitalization, prefixes and suffixes), The second line shows how including other types of information such as lemmas, POS, semantic tags (medical categories from the SNOMED-CT ontology) and Brown clusters and embeddings, the results improved considerably, reaching 70.30. Lines 3, 4 and 5 present the final results of our neural network leading to 72.01. We found that the unsupervised methods give encouraging improvements over the baseline. We made a final set of experiments trying to decide whether they are complementary or redundant. To that effect, we experimented with different combinations of pairs of unsupervised information types. The combinations of different embedding types did not give any significant improvement. All in all, the Brown clusters combined with the SkipNG embeddings gave the best results, presented in the last line of Table 3.

		MicroAvg.	Drug			Disease		
		F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
[38]	ne	53.80	84.80	54.92	66.67	60.23	39.27	47.55
	P+Bc+cE	70.30	**89.81**	84.90	87.29	66.22	57.14	61.69
Our work	ne	60.77	78.82	75.71	77.23	57.09	48.13	52.23
	SkipNG	70.60	87.70	85.78	86.73	66.67	58.48	62.31
	SkipNG+Bc	**72.01**	88.04	**88.62**	**88.33**	**66.86**	**60.73**	**63.65**

Table 3: Results of the system on the test set. Notation: ne=no embeddings; P=POS tags; Bc=Brown clusters; cE= clustered embeddings; SkipNG=Skip N-gram model.

5. Discussion

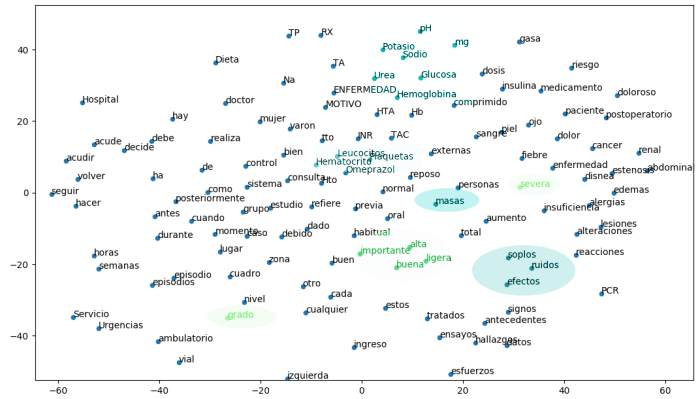
Results from Figure 2 revealed that using pretrained embeddings boosted the MER. Under all circumstances SkipNG outperformed the other algorithms indicating that word order determines the quality of the pretrained embeddings in a MER task. The fact that SkipNG uses an attention mechanism to calculate the embeddings, helps to select relevant words within the context to make more accurate predictions, which seems substantial for a sequential task such as NER. These results support work by Lample et al. [16] on generic NER, who mentioned this concern while showing no comparative results to sustain this conclusion.

[24] reports that the effect of hyperparameters should not be studied in isolation. Focusing on the application of a single algorithm, Figure 2 shows the domain and genre influence on the results, together with hyperparameters such as dimension and window size. Using EHRs (in-domain-genre data) clearly outperformed the other corpora in almost all of the embedding types and parameter settings. It is remarkable that differences between in-domain and out-domain are not as clear depending on the embeddings. Although from previous works [24], [26], it is difficult to reach any conclusion about the values of the best hyperparameters, Figure 2 shows that the best parameters are different for each tool, with the exception of SkipNG wich seems to be the most stationary in that respect.

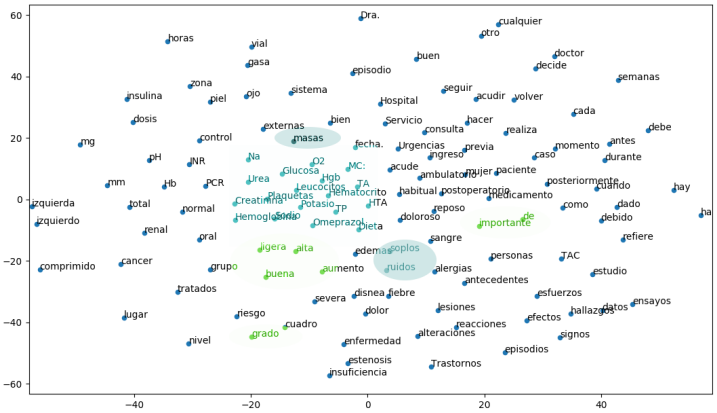
Regarding the importance of each factor, Figure 2 illustrates how embedding dimension and window are not as important as the corpus domain and genre. Therefore, choosing an appropriate algorithm should be the first step, then choosing the domain and corpus genre and, finally, the hyperparameters should be tuned.

On the other hand, the general Spanish corpus covers 56% of the words compared to 55% with the standard medical corpus, but the latter gives better results in general, with the exception of the GloVe embeddings (see Figure 2). This happens even when its size is much smaller (1.3M words compared to 1.5B), suggesting that the medical corpus gives better representations for the task at hand. The unsupervised information from EHRs (52M words) gives the best coverage on the training set (83%) and also the best results by a considerable difference. Hence, coverage is a sensitive indicator. Furthermore, when EHRs are used, the improvement of the results is motivated mainly by two factors: 1) the higher coverage on the training set and 2) the use of an in-domain corpus. As can be seen for GEN and GEN-MED corpora, we have obtained better results using an in-domain corpus than using a general domain corpus with similar coverage.

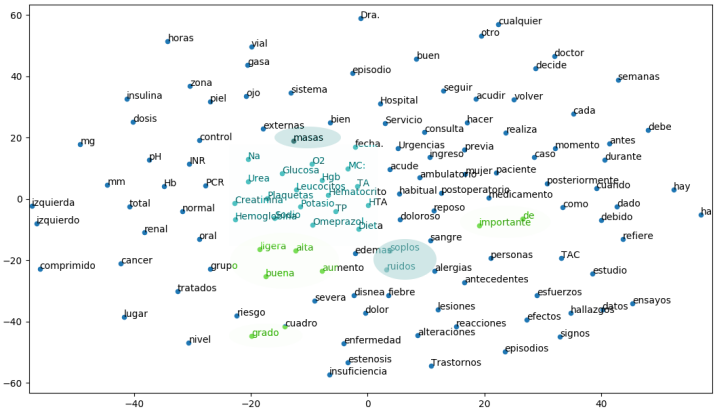
To illustrate the nature of the embeddings obtained from the General corpus, Medical corpus and EHRs, respectively, we selected the 150 most frequent words in the training set. Figures 4a, 4b and 4c show the word representations in a two dimensional space. We observe that similar entities such as blood substances, words in light blue *Potasio* (potassium), *Sodio* (sodium), or *Urea* (urea), are close to each other. The same occurs with some findings, words surrounded by a blue circle *masas* (masses), *soplos* (murmurs), *ruidos* (sounds) and some gradual modifiers, *importante* (important), *leve* (minor). Note that, while in the EHR embeddings these groups are coherent and well formed, in the Medical Corpus embeddings and General corpus embeddings several elements appear far from each other.



(a) General corpus



(b) Medical corpus



(c) EHRs

Figure 4: Bidimensional representation of the most frequent words in the training set obtained with dense representations from alternative sources.

From Table 2 we can conclude that character layer is useful in MER to detect prefixes and suffixes. The biggest improvement occurs when the character layer is used in association with the embeddings acquired from a corpora other than EHRs (GEN and GEN-MED). These results suggests that the use of the character layer in combination with a general domain corpus provides the system with the capacity to model the structure of the words in the language, alleviating the initial coverage problem due to the out-of-vocabulary words. On the other hand, the use of an in-domain corpus in combination with the character layer can lead the system to modelate the structure of the words of the language while it extends the specific vocabulary of the domain, obtaining the best improvements of the experimental setups.

The experiments presented in Figures 2 and 3 were performed on the development set. Table 3 presents the final results on the test set. Pérez et al. [38] required a feature engineering process establishing, for each knowledge type, a number of parameters such as window size, or combinations of different linguistic types of information (such as lemma, POS, or semantic tag). In row 3 we see how, using only word forms together with an embedding layer that is fed only on the training set, there is a boost in the F-score (60.77) compared to the simplest model of Pérez et al. [38]. Row 4 reveals that feeding external knowledge in the form of embeddings pretrained on large corpora gives a big improvement (70.60), outperforming the best result of the costly feature engineering approach (row 2 of Table 3).

As an example of the applicability of the resulting system, it has been also tested at the shared task on Disability Annotation on Documents from the Biomedical Domain[6], with the objective of detecting entities corresponding to disabilities (e.g., *mental retardation*, *blindness*) for English and Spanish. Our MER tagger obtained the best F-score for both languages: 82.1 and 78.6 for English and Spanish, respectively, 7.1 and 6.4 absolute points above the next best systems.

[6] http://nlp.uned.es/diann/

6. Conclusions

In the present work, we have evaluated neural models using three corpora and five different approaches to obtain word representations in an unsupervised manner, comparing a number of parameter settings. We have performed a deep study of their combinations so as to measure joint interactions. The resulting deep learning model gives a significant improvement on the state-of-the-art in Medical Named Entity Recognition of diseases and drugs in Spanish EHRs. We have shown that the performance of our Bi-LSTM+CRF based MER system on EHRs improves using pretrained word dense representations and how the algorithm, the domain and genre of the corpus, and the hyperparameter setting employed to obtain these representations have a big impact on the results. It is crucial to choose the right algorithm to learn the dense word representations. Our study shows that SkipNG, which takes word order into account, outperforms other typically employed algorithms such as GloVe or Word2Vec. A relevant conclusion is that certain word representations proved to be complementary and therefore using them ensembled improved the results as in the case of Brown clusters and SkipNG word embeddings. Secondly, in the case of EHRs, training the embeddings on the same corpus domain and genre is very important as well. However, obtaining freely available EHRs is not an easy enterprise and the results revealed that learning from a medical general corpus is more convenient than using a 1,000 times larger general out-domain corpus. We have also shown, when training with EHRs is not possible, how using a character embedding layer allows us to capture prefix and suffix information improving the results and how it helps to reduce the gap between the EHR embeddings and medical or out-domain corpus embeddings.

Our findings offer a guide to help decide among different possibilities and strategies for MER, showing that, in several cases, using the default paramete rvalues or most popular tools without further exploration of alternative settings is far from optimal. We hope that the present work serves as a guide to other researchers who envisage doing MER, proposing alternatives to using the

standard settings.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Economy, Industry and Competitiveness (PROSA-MED: TIN2016- 77020-C3-1-R, TUNER: TIN2015-65308-C5-1-R) and the Basque Government (BERBAOLA: KK-2017/00043).

References

[1] D. Clifton, K. Niehaus, P. Charlton, G. Colopy, Health informatics via machine learning for the clinical management of patients, Yearbook of medical informatics 10 (1) (2015) 38–43.

[2] W. Ning, M. Yu, Exploiting distributional semantics to benefit machine learning in automated classification of Chinese clinical text, in: Int. Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 1096–1102, 2016.

[3] K. Jensen, C. Soguero-Ruiz, K. O. Mikalsen, R.-O. Lindsetmo, I. Kouskoumvekaki, M. Girolami, S. O. Skrovseth, K. M. Augestad, Analysis of free text in electronic health records for identification of cancer patient trajectories, Scientific Reports 7 (2017) https://doi.org/10.1038/srep46226.

[4] F. Doshi-Velez, Y. Ge, I. Kohane, Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis, Pediatrics 133 (1) (2014) 54–63.

[5] P. Schulam, F. Wigley, S. Saria, Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery., in: AAAI, 2956–2964, 2015.

[6] T. A. Lasko, J. C. Denny, M. A. Levy, Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data, PloS one 8 (6) (2013) https://doi.org/10.1371/journal.pone.0066341.

[7] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, G. Gonzalez, Utilizing social media data for pharmacovigilance: A review, Journal of biomedical informatics 54 (2015) 202–212.

[8] D. L. Mowery, B. R. South, L. Christensen, J. Leng, L.-M. Peltonen, S. Salanterä, H. Suominen, D. Martinez, S. Velupillai, N. Elhadad, et al., Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2, Journal of biomedical semantics 7 (1) (2016) 43.

[9] R. Leaman, R. Khare, Z. Lu, Challenges in clinical natural language processing for automated disorder normalization, Journal of biomedical informatics 57 (2015) 28–37.

[10] A. Pérez, A. Atutxa, A. Casillas, K. Gojenola, Á. Sellart, Inferred joint multigram models for medical term normalization according to ICD, International journal of medical informatics 110 (2018) 111–117.

[11] S. Santiso, A. Perez, A. Casillas, Exploring Joint AB-LSTM with embedded lemmas for Adverse Drug Reaction discovery, IEEE journal of biomedical and health informatics .

[12] W. H. Organization, Management of patient information: Trends and challenges in Member States, Global Observatory for eHealth series 6.

[13] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: Proceedings of the seventh conference on Natural language learning NAACL 2003-Volume 4, ACL, 1–4, 2003.

[14] J. Kazama, T. Makino, Y. Ohta, J. Tsujii, Tuning support vector machines for biomedical named entity recognition, in: Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3, Association for Computational Linguistics, 1–8, 2002.

[15] B. Strauss, B. Toma, A. Ritter, M.-C. de Marneffe, W. Xu, Results of the wnut16 named entity recognition shared task, in: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), 138–144, 2016.

[16] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition., in: K. Knight, A. Nenkova, O. Rambow (Eds.), HLT-NAACL, The Association for Computational Linguistics, ISBN 978-1-941643-91-4, 260–270, 2016.

[17] M. Rei, Semi-supervised Multitask Learning for Sequence Labeling, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2121–2130, 2017.

[18] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, et al., Relation Classification via Convolutional Deep Neural Network., in: COLING, 2335–2344, 2014.

[19] J. Schmidhuber, Deep learning in neural networks: An overview, Neural networks 61 (2015) 85–117.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 3111–3119, 2013.

[21] K. Roberts, Assessing the corpus size vs. similarity trade-off for word embeddings in clinical NLP, in: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), 54–63, 2016.

[22] S. Lai, K. Liu, S. He, J. Zhao, How to generate a good word embedding, IEEE Intelligent Systems 31 (6) (2016) 5–14.

[23] O. Melamud, D. McClosky, S. Patwardhan, M. Bansal, The Role of Context Types and Dimensionality in Learning Word Embeddings, in: NAACL HLT 2016, 1030–1040, 2016.

[24] B. Chiu, A. Korhonen, S. Pyysalo, Intrinsic evaluation of word vectors fails to predict extrinsic performance, in: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, 1–6, 2016.

[25] A. N. Jagannatha, H. Yu, Bidirectional rnn for medical event detection in electronic health records, in: Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting, vol. 2016, NIH Public Access, 473, 2016.

[26] A. N. Jagannatha, H. Yu, Structured prediction models for RNN based sequence labeling in clinical text, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2016, 856, 2016.

[27] K. Xu, Z. Zhou, T. Hao, W. Liu, A Bidirectional LSTM and Conditional Random Fields Approach to Medical Named Entity Recognition, in: International Conference on Advanced Intelligent Systems and Informatics, Springer, 355–365, 2017.

[28] P. Stenetorp, H. Soyer, S. Pyysalo, S. Ananiadou, T. Chikayama, Size (and domain) matters: Evaluating semantic word space representations for biomedical text, Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM), 2012 .

[29] M. Oronoz, A. Casillas, K. G. eta Alicia Prez, Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names, in: Lecture Notes in Computer Science, 8259. Progress in Pattern Recognition, ImageAnalysis, ComputerVision, and Applications 18th Iberoamerican Congress, CIARP 2013 Havana, Cuba, November 20-23, 2013 Proceedings, Part II, 536–543, 2013.

[30] L. Padr, E. Stanilovsky, FreeLing 3.0: Towards Wider Multilinguality, in: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA, Istanbul, Turkey, 2473–2479, 2012.

[31] M. Oronoz, K. Gojenola, A. Prez, A. D. de Ilarraza, A. Casillas, On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions, Journal of Biomedical Informatics 56 (2015) 318–332.

[32] S. Pradhan, N. Elhadad, B. R. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W. W. Chapman, G. Savova, Evaluating the state of the art in disorder recognition and normalization of the clinical narrative, Journal of the American Medical Informatics Association 22 (1) (2014) 143–154.

[33] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[34] J. A. Bullinaria, J. P. Levy, Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD, Behavior research methods 44 (3) (2012) 890–907.

[35] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, J. C. Lai, Class-based n-gram models of natural language, Computational linguistics 18 (4) (1992) 467–479.

[36] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532–1543, 2014.

[37] W. Ling, Y. Tsvetkov, S. Amir, R. Fermandez, C. Dyer, A. W. Black, I. Trancoso, C.-C. Lin, Not All Contexts Are Created Equal: Better Word Representations with Variable Attention., in: L. Mrquez, C. Callison-Burch, J. Su, D. Pighin, Y. Marton (Eds.), EMNLP, The Association for Computational Linguistics, 1367–1372, 2015.

[38] A. Pérez, R. Weegar, A. Casillas, K. Gojenola, M. Oronoz, H. Dalianis, Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora, Journal of Biomedical Informatics 71 (2017) 16–30.

Summary points

What was already known on the topic?

- Most Medical Entity Recognition (MER) systems use SVM, Perceptron or CRF based machine learning techniques. These approaches require heavy feature engineering and preprocessing.
- Recently, neural network approaches have being proposed, simplifying the recognition task because no preprocessing is required, but they are highly dependent on the quality of the word representations they use.

What does this work add?

- We adapted the Bi-LSTM proposed by Lample et al. [16] to make it encompass multiple dense representations as well as POS and semantic tags in parallel.
- We evaluated neural models using three corpora and five different word representation approaches in an unsupervised manner, comparing a number of parameter settings.
- We have performed a deep study of their combinations, so as to measure joint interactions.
- Our results offer a guide to help decide among different possibilities and strategies for MER, showing that in several cases using the standard settings or most popular tools without further exploration of alternative settings is far from optimal.