Medical Word Embeddings for Spanish: Development and Evaluation

Felipe Soares
Barcelona Supercomputing Center (BSC)
felipe.soares@bsc.es

Marta Villegas
Barcelona Supercomputing Center (BSC)
marta.villegas@bsc.es

Aitor Gonzalez-Agirre
Barcelona Supercomputing Center (BSC)
aitor.gonzalez@bsc.es

Martin Krallinger
Centro Nacional de Investigaciones Oncologicas (CNIO)
mkrallinger@cnio.es

Jordi Armengol-Estapé
Universitat Politècnica de Catalunya (UPC)
jordi.armengol.estape@gmail.com

Abstract

Word embeddings are representations of words in a dense vector space. Although they are not recent phenomena in Natural Language Processing (NLP), they have gained momentum after the recent developments of neural methods and Word2Vec. Regarding their applications in medical and clinical NLP, they are invaluable resources when training in-domain named entity recognition systems, classifiers or taggers, for instance. Thus, the development of tailored word embeddings for medical NLP is of great interest. However, we identified a gap in the literature which we aim to fill in this paper: the availability of embeddings for medical NLP in Spanish, as well as a standardized form of intrinsic evaluation. Since most work has been done for English, some established datasets for intrinsic evaluation are already available. In this paper, we show the steps we employed to adapt such datasets for the first time to Spanish, of particular relevance due to the considerable volume of EHRs in this language, as well as the creation of in-domain medical word embeddings for the Spanish using the state-of-the-art FastText model. We performed intrinsic evaluation with our adapted datasets, as well as extrinsic evaluation with a named entity recognition systems using a baseline embedding of general-domain. Both experiments proved that our embeddings are suitable for use in medical NLP in the Spanish language, and are more accurate than general-domain ones.

1 Introduction

Representation of words in vector space, or word embedding, is not a new concept in Natural Language Processing (NLP) and are used in a several number of statistical and neural models (Ghannay et al., 2016). Word embeddings (WE) can include semantic information and are based on the general idea of an association of elements (words) with certain contexts and the similarity in word meanings. In more recent neural networks, embeddings are used to encode words in a space that is subsequently used as input for many possible models.

1.1 Background

In the work of Mikolov et al. (2013a), they introduced two new architectures for estimating continuous representations of words using log-linear models, called continuous bag-of-word (CBOW) and continuous skip-gram (skip-gram). CBOW calculates the projection for the current word given the context words in the particular sentence, while skip-gram, following its name, skip the word being processed and evaluates projections of the context words. Further works gave more insights about this method called Word2Vec (Mikolov et al., 2013b,c). Since its appearance, Word2Vec has been used and adapted for a wide range of applications, including sentiment analysis (Nakov et al., 2016; Yu et al., 2017), named entity recognition (Chiu and Nichols, 2016), clas-

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 124–133
Minneapolis, Minnesota, June 7, 2019. ©2019 Association for Computational Linguistics

sification (Zhang et al., 2015), clustering (Kim et al., 2017), word sense disambiguation (Iacobacci et al., 2016) and many others. More recently, Mikolov et al. (2018) presented the combination of various ”tricks” in training word embeddings that are rarely used together, but that outperforms the previous state-of-the-art vector representations.

1.2 Pre-trained embeddings

Pre-trained word embeddings are widely available for a plethora of languages and methods. Google, for instance, makes available Word2Vec models pre-trained on about 100 billion words from Google News corpus in English[1]. Regarding other languages, on FastText website[2] one can download pre-trained embeddings for 157 languages based on Common Crawl and Wikipedia. For the specific case of Spanish, the University of Chile NLP group makes available FasText and Word2Vec embeddings[3] using the Spanish Billion Word Corpus (SBWCE)[4].

1.3 Biomedical embeddings

As pointed out by Chiu et al. (2016), most of the studies and available embeddings are focused on general-domain texts and general evaluation datasets. Thus, their results not necessarily apply well to medical and biomedical text analysis. Their study, in English, demonstrates that bigger corpora do not necessarily produce better biomedical word embeddings. They also made their resulting embeddings available for download.

In another work, Chen et al. (2018) created sentence embeddings for clinical and biomedical texts, called BioSentVec trained on PubMed and clinical notes from the MIMIC-III Clinical Database(Johnson et al., 2016). Similarly, Sahu and Anand (2015) used the PubMed Central Open Access subset (PMC) and PubMed abstracts to train word embeddings for English using CBOW. They evaluate embeddings performance using similarity and relatedness datasets, which will be presented in Section 3.1. However, they do not compare the trained models with a general-domain one.

In a more fine-grained application, Zhang et al. (2018) adapted word embeddings to recognize symptoms in the target domain of psychiatry. As a source for their embeddings, they used four corpora: intensive care, biomedical literature, Wikipedia and Psychiatric Forum. Ling et al. (2017) developed a method to integrate extra knowledge into word embeddings for biomedical NLP tasks via graph regularization.

More related to our work, Santiso et al. (2018) developed word embeddings tailored for negation detection in health records written in Spanish. As corpora, they used both in-domain and general-domain data. For in-domain, they used unannotated Electronic Health Records (EHRs) from a hospital in Spain. For the general-domain, they used the SBWCE corpus. However, they did not perform any intrinsic evaluation of the generated embeddings; neither made them available for use or compared general-domain and in-domain performance.

Also regarding Spanish biomedical embeddings, the work of Segura-Bedmar and Martínez (2017) shows the use of pre-trained word embeddings with SBWCE for simplification of drug package leaflets so that they are more friendly to the patients. However, they do not use in-domain embeddings for such task. Also, Villegas et al. (2018) collected a census of Spanish texts that can be of use in text mining, however, they did not provide any sort of word embeddings.

1.4 Contributions and Structure

Given that very little attention has been given to producing and evaluating quality word embeddings in Spanish for the biomedical domain, we propose to develop embeddings based on the state-of-the-art FastText model with in-domain data. In addition, only works aiming the English language provide a comprehensive performance evaluation of in-domain embeddings when compared to general-domain ones. For that, we will adapt them to Spanish. We claim as relevant the following contributions:

- Development of Spanish embeddings for the Biomedical domain;

- Intrinsic and extrinsic evaluation of performance using established datasets and a Named Entity Recognition (NER) task;

[1] https://code.google.com/archive/p/word2vec/

[2] https://fasttext.cc/docs/en/crawl-vectors.html

[3] https://github.com/uchile-nlp/spanish-word-embeddings

[4] http://crscardellino.github.io/SBWCE/

- Comparison of in-domain and general-domain performance;

- Adaptation of established biomedical intrinsic evaluation datasets for the Spanish language;

- Embeddings are public available[5] and licensed under CC-BY 4.

We expect that the developed word embeddings will to be used in several clinical NLP applications, such as for the identification of sections in clinical documents since the embedddings can be used to create phrase and paragraph embeddings. Also, for text summarization based on neural networks, our embeddings can be used as a resource during training.

The rest of the paper is organized as follows. In Section 2, we explain the methods and the materials used in our experiments, including corpora and the training procedure. In Section 3, we detail the intrinsic and extrinsic evaluations, with the steps we employed to adapt English datasets to Spanish. In Section 4, we show the experiments and their results, while in Section 5 we perform a brief discussion and conclusion.

2 Material and Methods

In this Section, we present the corpora, the word embedding model used in our study and the training procedure.

2.1 FastText

The FastText model (Mikolov et al., 2018) uses the 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 subword information in a combined manner. The CBOW and skip-gram models is the same as proposed in Mikolov et al. (2013a).

The position-dependent weighting introduces information regarding the position of the word being evaluated. As stated by the authors, the explicit encoding of the word and its position would lead to overfitting. The solution was to learn position representations and use them to reweight the word vectors at a minimum computational cost using linear combination of both representations.

The original Word2Vec is insensitive to word order, since it is only based on unigrams. To capture word order information in a phrase representation, the authors merge words with high mutual information in a single token. One example can be ”brain” and ”dead”, which could be merged as ”brain_dead”. This process of merging tokens can be repeated several times to produce longer tokens.

To avoid the fact that standard word vectors ignore word-internal structure, which may contain useful information, the authors enrich the vectors with subword information. Each word is decomposed into its character n-grams which are then learned. After that, the final word vector is the simple sum of the word vector and their n-grams representations.

2.2 Corpora

To develop our in-domain embeddings, we used two sources of data: (i) the SciELO database, which contains full-text articles primarily in English, Spanish and Portuguese, and (ii) the Wikipedia, with a subset which we call Wikipedia Health, comprised by the categories of Pharmacology, Pharmacy, Medicine and Biology. This method of combining large corpora (i.e. SciELO) and smaller focused (i.e. Wikipedia) was shown to be an adequate approach to produce quality embeddings for clinical NLP (Roberts, 2016). The choice of SciELO is that this database is the most comprehensive in term of number of articles and abstracts available in Spanish. As for the Wikipedia, it can be a source of information for specific terms, which can benefit our models.

From Scielo.org, all documents in Spanish were downloaded, language checked and processed into sentences. For language check, we used the langdetect library[6] for Python. The scielo.org node contains all Spanish articles, regardless if they are from European or Latin American Spanish. In the database, articles from the health domain correspond to approximately 50% of the results.

Using the Wikipedia API for Python[7], we retrieved all articles that are from the aforementioned categories. We also performed language checking, to ensure that all sentences were in Spanish.

[5] http://doi.org/10.5281/zenodo.2542722

[6] https://github.com/fedelopez77/langdetect

[7] https://pypi.org/project/wikipedia/

In Table 1, one can see the statistics regarding the gathered corpora. Sentences were produced using the sentence tokenizer from the NLTK package. The SciELO corpus is relatively smaller than the Wikipedia one regarding number of sentences. However, as for number of tokens, SciELO contains almost 22% more than Wikipedia. This is probably due to the fact that scientific article sentences are longer than the ones available in Wikipedia.

Table 1: Statistics for the gathered corpora

Corpus	Sentences	Tokens
SciELO Full-Text	3.3M	100M
Wikipedia Health	4M	82M

2.3 Training

We used the FastText implementation available in `https://fasttext.cc` to train our word embeddings. The following setup was used:

- Minimum number of word occurrences: 5
- Phrase representation: No (i.e. length of word n-gram = 1)
- Minimum length of character n-gram: 3
- Maximum length of character n-gram: 6
- Size of word vectors: 300
- Epochs: 20

3 Evaluation

For the evaluation of our embeddings, we use both intrinsic and extrinsic evaluation, which are now detailed, as well as the baseline word embedding.

3.1 Intrinsic

In the intrinsic evaluation, the performances are measured regarding specific tasks that are only related to the embedding itself, such as syntactic of semantic relationships between words. The most common examples are similarity, relatedness and analogy evaluations (Schnabel et al., 2015).

For the biomedical domain, some standard datasets are available for the evaluation of semantic similarity and relatedness. The UMNSRS similarity (UMNSRS-sim) and UMNSRS relatedness (UMNSRS-rel) are datasets consisting of pairs of UMLS (Unified Medical Language System) concepts manually annotated for similarity and relatedness. Details about the original datasets can be found in Pakhomov et al. (2010). The UMNSRS-sim contains 566 pairs of concepts, while the UMNSRS-rel contains 587 pairs.

Another well-known dataset for intrinsic evaluation in biomedical embeddings is the MayoSRS (Pakhomov et al., 2011), which is used for similarity evaluation and is comprised of 101 UMLS pairs and their respective manual scores.

The aforementioned datasets, however, are only available in English. For the best of our knowledge, no standard Spanish dataset is available for the biomedical domain. Thus, in order to be able to evaluate our embeddings, we adapted the aforementioned datasets for Spanish.

In Figure 1, we depict the steps employed to adapt the datasets. In step 1, the datasets are translated to Spanish using Google Translate[8]. However, due to the possible polysemy and translation errors, we employed additional checking steps.

In step 2, the translated terms are queried against the already available translations for that specific CUI (Concept Unique Identifier) in UMLS. If the translated term is already in the UMLS translations, we assign such term as a valid translation.

In step 3, if the translated term is not found in UMLS, we perform manual evaluation of possible translations using UMLS browser. The assigned translations were then revised by a medical doctor and corrected when needed. Also, at this point all other assignd terms were also revised.

We must notice that we did not include the concepts that were originally referring to commercial drug names (which are not in the UMLS, just their pharmacological substance), since this may vary depending on the country and also depending on regional medical protocols. The final number of pairs of terms for UMNSRS-rel is 384, that is, 65.41% of the original in English. As for UMNSRS-sim, the final number is 380, or 67.14% of the original dataset in English. For the MayoSRS, all 101 pairs are included in the final dataset in Spanish, since no drug is included in the original data.

[8] `https://translate.google.com/`

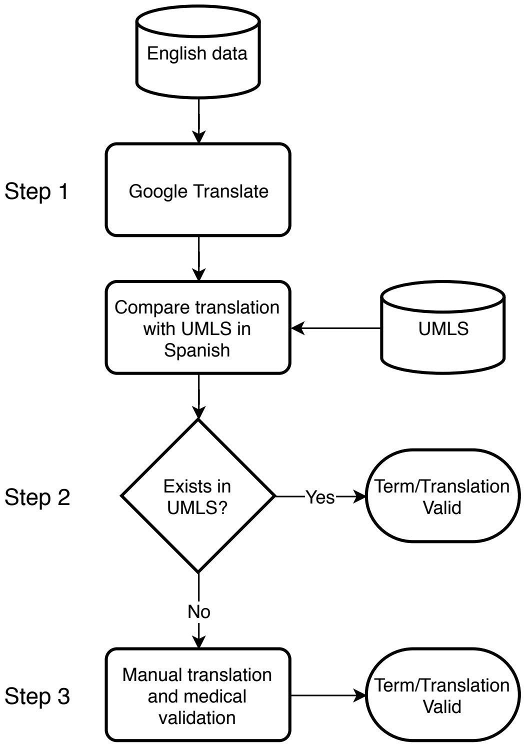

Figure 1: Steps performed to translate the UMNSRS-sim, UMNSRS-rel and MayoSRS datasets to Spanish

3.2 Extrinsic

As for the extrinsic evaluation, we employed our embeddings in an NER task to identify pharmacological substances, compounds and proteins in clinical texts.

3.2.1 Data

The data for this experiment comprehends manually classified collection of clinical case sections derived from Open access Spanish medical publications, named the Spanish Clinical Case Corpus (SPACCC). All clinical case records derived from various databases were gathered in a first step, preprocessed and the actual clinical case section was extracted removing embedded figure references or citations. These records where classified manually using the MyMiner[9] file labeling online application by a practicing oncologist and revised by a clinical documentalist in order to assure that these records were related to the medical domain and they resembled the kind of structure and content that is relevant to process clinical content.

The final collection of 1000 clinical cases that make up the corpus had a total of 16504 sentences, with an average of 16.5 sentences per clinical case. The SPACCC corpus contains a total of 396,988 words, with an average of 396.2 words per clinical case. It is noteworthy to say that this kind of narrative shows properties of both, the biomedical and medical literature as well as clinical records. Moreover, the clinical cases were not restricted to a single medical discipline, and thus cover a variety of medical topics, including oncology, urology, cardiology, pneumology or infectious diseases, which is key in order to cover a diverse collection of chemicals and drugs.

We must notice that this corpus will not be available at this point since it is currently being used as evaluation in a shared task track. However, in the future, users will be able to access the corpus from the same link to the word embeddings.

3.2.2 Software

As for the NER system, we employed an off-the-shelf framework called NeuroNER(Dernoncourt et al., 2017)[10]. The engine is based on artificial neural networks, relying on long short-term memory (LSTM) to predict the label of a sequence of tokens. The network contains three main layers: (i) the character-enhanced token-embedding layer, (ii) the label prediction layer, and (iii) the label sequence optimization layer. The word embeddings are fed to the first layer (i.e token-embedding).

3.2.3 Baseline Word Embedding

As a baseline for our comparisons, we decided to use the embeddings available from the University of Chile NLP Group[11]. The embeddings are trained based on the SBWC corpus and the training settings are the same we have shown in Section 2.3, thus making our comparisons fair.

One big difference between our training process is related to the corpora used. SBWC is a general-domain corpus, comprised of approximately 1.4 billion words, while our combined corpora contain roughly 1.2 million words. Thus, the general-domain corpus is approximately one order of magnitude larger than ours.

4 Experiments and Results

In this section, we detail how the experiments were carried out and the results we obtained for both intrinsic and extrinsic evaluation methods, as

[9] http://myminer.armi.monash.edu.au

[10] http://neuroner.com/

[11] https://github.com/uchile-nlp/spanish-word-embeddings

well as the comparisons with the baseline embedding presented in Section 3.2.3 and our embeddings, which we now call Spanish Health Embedding (SHE).

4.1 Intrinsic

In the intrinsic experiment, for the sake of a fair comparison between our proposed embedding and the baseline, we made sure that all the pairs being compared were available both in SHE and in the SBWC. For this, we checked for each pair of translated CUIs (explained in Section 3.1) if the words were present in both embeddings vocabularies. For multi-word terms, we averaged individual word vectors to compose the final term vector. The final number of compared pairs for each translated dataset are: UMNSRS-sim(322), UMNSRS-rel(252) and MayoSRS(101).

Regarding the evaluation, we calculated the cosine distance for each pair of terms and later compared those values with the human annotated ones in the datasets by means of Pearson correlation coefficient (ρ).

In Table 2, we depict the results for the comparison for each dataset regarding the Pearson correlation coefficient. One can notice that SHE presented the highest coefficient for the three used datasets by a large margin, being such statistically significant for all of them, except to SBWC with the MayoSRS dataset. Thus, as for intrinsic evaluation, we can assume that our embeddings are better than the general-domain embedding trained on SBWC.

Table 2: Comparison of the intrinsic evaluation between the proposed embeddings (SHE) and the general-domain ones (SBWC). Bold numbers represent the best results for each dataset, while asterisc means that such coefficient was statistically significant.

	SHE (our)	**SBWC**
Dataset	ρ	ρ
UMNSRS-sim	**0.5826***	0.4319*
UMNSRS-rel	**0.5239***	0.3947*
MayoSRS	**0.3174***	0.1237

4.2 Extrinsic

For the extrinsic evaluation, we used the NeuroNER framework, which was described in Section 3.2.2, with a biomedical corpus of clinical notes described in Section 3.2.1. The corpus has 4 entity labels: Proteins, Normalizable Chemicals, No-Normalizable Chemicals, and Unclear mentions. The reason for such lables is that they can be normalized to a fixed ontology, in the case of Proteins and Chemicals, while some chemicals cannot be normalzied or are unclear. Since the number of ”No-Normalizable” mentions is very low compared to all labels, we did not include them in our evaluation.

We trained NeuroNER with the following standard parameters using our embeddings and the SBWC one:

- Data splitting: 80% training, 10% validation, 10% test. Stratified and fixed for both embeddings;
- Character-embedding dimension: 25
- Charater LSTM hidden state dimension : 25
- Token LSTM hidden state dimension: 300
- Patience: 10
- Maximum number of epochs: 100
- Optimizer: SGD
- Learning rate: 0.005
- Dropout rate: 0.5

In Table 3 we show the results of our embeddings compared to the SBWC trained with the same parameters as detailed in Section 2.3. One can notice that our proposed embedding achieved the best results in the validation set for all the named entity labels. As for the test set, we achieved the best scores in 8 out of 13 possible evaluations. But we must notice that as overall performance, our system achieved an F1 score of 88.18%, while the baseline achieved only 87.76%. Thus, our embeddings showed to be superior to general-domain one in this extrinsic evaluation.

4.3 Visual Evaluation

In Figures 2 and 3, we show the PCA (Principal Component Analysis) projections of our embeddings and the SBWC, respectively. We tried to follow the standards of Pakhomov et al. (2010) to categorize the terms using UMLS semantic types in the following categories: symptoms, diseases and drugs. Better quality and larger figures can be accessed online[12]

[12] http://doi.org/10.5281/zenodo.2542722

Table 3: Comparison of the extrinsic evaluation between the proposed embeddings (SHE) and the general-domain ones (SBWC). Bold numbers represent the best results for each metric and data parition, with Val meaning validation set.

	SHE (our)		**SBWC**	
	Val	Test	Val	Test
Overall				
Accuracy	**99.51**	**99.62**	99.45	99.57
Precision	**90.63**	90.42	90.30	**90.87**
Recall	**88.25**	**86.03**	86.12	84.45
F1	**89.42**	**88.17**	88.16	87.76
Normalizables				
Precision	**92.82**	93.18	91.87	**93.93**
Recall	**89.81**	88.09	88.89	**88.34**
F1	**91.29**	90.56	90.35	**91.05**
Proteins				
Precision	**87.86**	**86.94**	88.22	86.19
Recall	**87.86**	**84.52**	84.39	81.75
F1	**87.86**	**85.71**	86.26	83.91
Unclear				
Precision	**100**	84.21	92.86	**88.24**
Recall	**81.25**	**84.21**	81.25	78.95
F1	**89.66**	**84.21**	86.67	83.33

One can notice that in Figure 2, there is some overlapping between the disease and symptoms categories, but they are not as much overlapped as shown in Figure 3. In addition, in our embeddings, on the top of the drugs cluster, one can see that most of the antibiotics are clustered together (e.g. *penicilina*, *eritromicina*, *cefazolina*, *doxiciclina*). However, in the SBWC projection, such drugs are spread inside the cluster. Interestingly, for both embeddings, the words *hierro*, *calamina*, *ajo*, *alcohol* are the ones that are more closer to the other two clusters.

5 Discussion and Conclusion

By the intrinsic and extrinsic experiments performed in Sections 4.1 and 4.2 we were able to show that our proposed embeddings can provide better performance than a general-domain one, even being trained in a corpus one order of magnitude smaller. We made our embeddings available in http://doi.org/10.5281/zenodo.2542722.

By performing a visual evaluation of the PCA projections of our embeddings and a general-domain one, we also provided strong evidence that the ones trained in a in-domain corpus can provide better-defined clusters of words.

We oversee that the embeddings we provide can be used in many different applications that require them as a resource, especially the ones which employ artificial neural networks. For instance, we studied the application in a named entity recognition example, but they can be used for sentence similarity evaluation, text classification, machine translation, clustering, relation extraction, for instance.

6 Acknowledgements

This work was supported by the Encargo de Gestion SEAD-CNIO and SEAD-BSC of the Spanish National Plan for the Advancement of Language technologies, the ICTUSnet INTERREG Sudoe programme and the Amazon AWS Cloud Credits for Research.

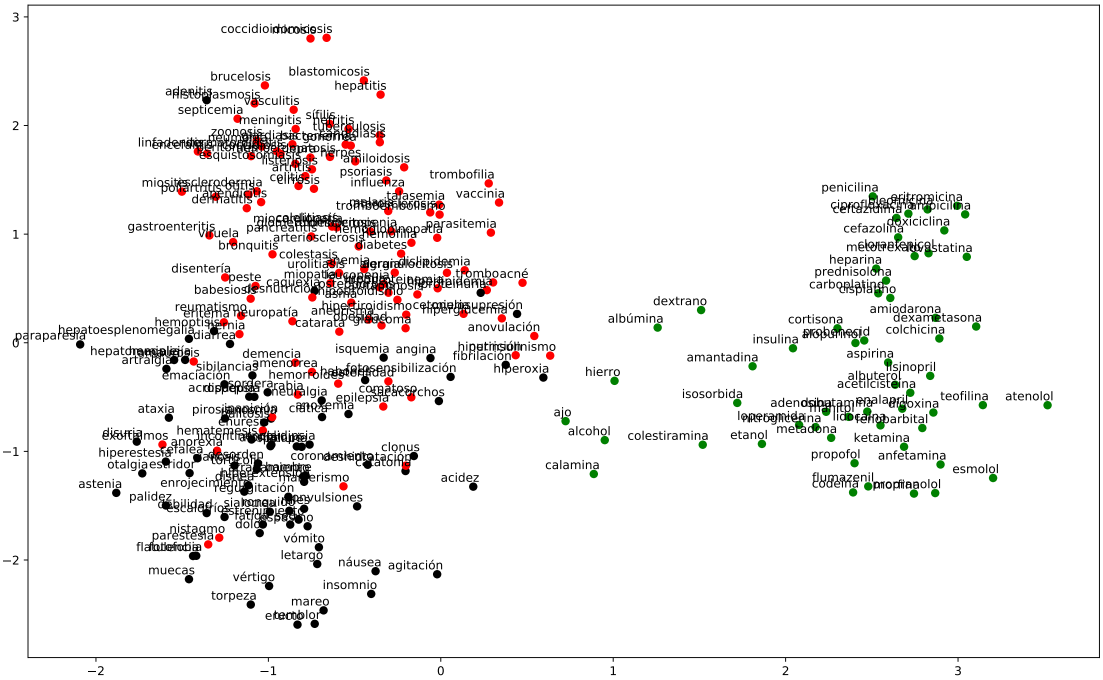

Figure 2: PCA projection of the UMNSRS concepts using our embeddings. Black means symptoms-related terms, red means disease-related terms, while green means drug-related terms.

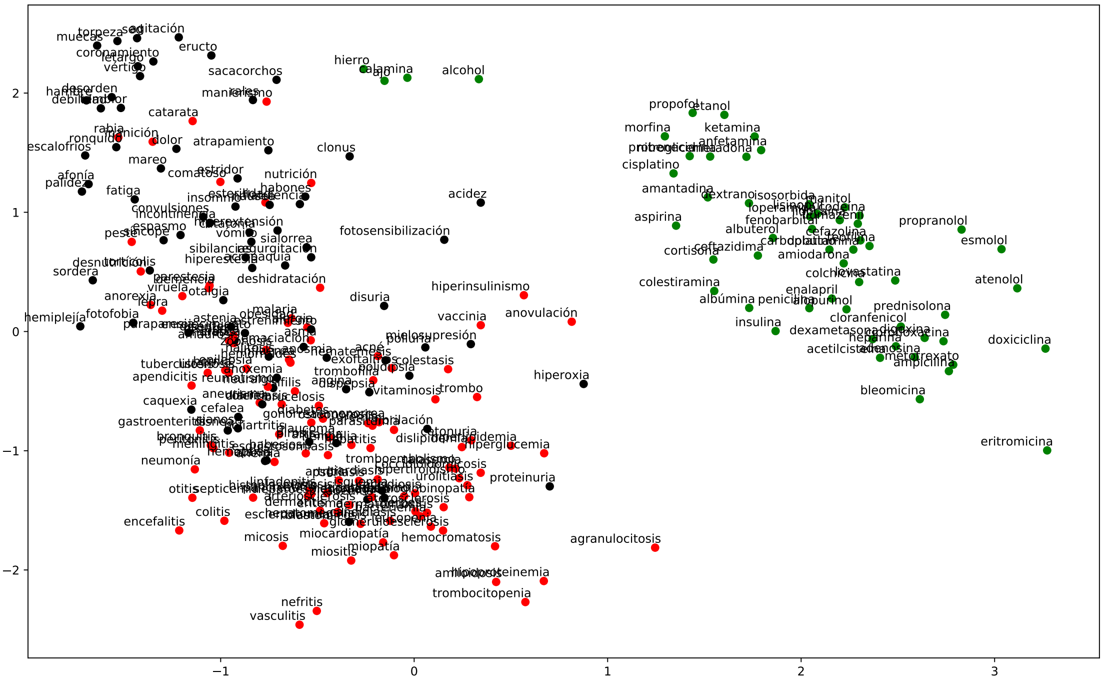

Figure 3: PCA projection of the UMNSRS concepts using the SBWC embeddings. Black means symptoms-related terms, red means disease-related terms, while green means drug-related terms.

References

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2018. Biosentvec: creating sentence embeddings for biomedical texts. *arXiv preprint arXiv:1810.09302*.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neuroner: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*.

Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. 2016. Word embedding evaluation and combination. In *LREC*, pages 300–305.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 897–907.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Han Kyul Kim, Hyunjoong Kim, and Sungzoon Cho. 2017. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352.

Yuan Ling, Yuan An, Mengwen Liu, Sadid A Hasan, Yetian Fan, and Xiaohua Hu. 2017. Integrating extra knowledge into word embedding models for biomedical nlp tasks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 968–975. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pages 1–18.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings*, volume 2010, page 572. American Medical Informatics Association.

Serguei VS Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B Melton, Alexander Ruggieri, and Christopher G Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics*, 44(2):251–265.

Kirk Roberts. 2016. Assessing the corpus size vs. similarity trade-off for word embeddings in clinical nlp. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 54–63.

Sunil Sahu and Ashish Anand. 2015. Evaluating distributed word representations for capturing semantics of biomedical concepts. *Proceedings of BioNLP 15*, pages 158–163.

Sara Santiso, Arantza Casillas, Alicia Pérez, and Maite Oronoz. 2018. Word embeddings for negation detection in health records written in spanish. *Soft Computing*, pages 1–7.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Isabel Segura-Bedmar and Paloma Martínez. 2017. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of biomedical semantics*, 8(1):45.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimón, and Martin Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. In *Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing"*, Paris, France. European Language Resources Association (ELRA).

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539.

Dongwen Zhang, Hua Xu, Zengcai Su, and Yunfeng Xu. 2015. Chinese comments sentiment classification based on word2vec and svmperf. *Expert Systems with Applications*, 42(4):1857–1863.

Yaoyun Zhang, Hee-Jin Li, Jingqi Wang, Trevor Cohen, Kirk Roberts, and Hua Xu. 2018. Adapting word embeddings from multiple domains to symptom recognition from psychiatric notes. *AMIA Summits on Translational Science Proceedings*, 2017:281.