Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 37-44 recibido 01-05-2021 revisado 06-06-2021 aceptado 08-06-2021

Constructing Corpus and Word Embedding for Spanish Covid-19 Data

Construcción de corpus y word embedding para datos de Covid-19 en español

Kyungjin Hwang
Korea University, Seoul, Republic of Korea
kjhwang0624@korea.ac.kr

Abstract Severe acute respiratory syndrome coronavirus 2 (COVID 19), colloquially referred to as coronavirus, escalated into a global pandemic with severe transmission and mortality rates in 2019. Despite the escalation of the virus' worldwide impact in 2020, numerous studies on Natural Language Processing in Spanish have neglected corpus construction or word embedding, especially conspicuous in its absence being the corpora involving coronavirus or infectious diseases. Additionally, corpus construction or word embedding conducted in the medical field do not display efficacy in production pertaining to coronavirus or infectious diseases. To supplement this potentially detrimental insufficiency, this study collects Spanish Language data to build a relevant coronavirus corpus through appropriate preprocessing and then obtains a word embedding. Performance of the corpus and word embedding are then tested through word similarity evaluations, a cosine similarity evaluation, and a visualization evaluation with the existing Spanish corpus. After comparison, corpus and word embedding suitable for coronavirus will be suggested.
Keywords: corpus, word embedding, coronavirus.

Resumen La Enfermedad Infecciosa por Coronavirus-19 (en adelante Covid-19), que comenzó a extenderse globalmente en diciembre de 2019, mostró una alta tasa de infección y mortalidad, y tuvo un gran impacto en el mundo en 2020. Sin embargo, los estudios existentes de procesamiento del lenguaje natural en español no han utilizado la construcción de corpus o la incrustación de palabras para enfermedades infecciosas, incluido el coronavirus. La construcción de corpus y la incrustación de palabras en el campo biomédico no han mostrado un rendimiento eficaz en la ayuda para luchar contra las enfermedades infecciosas, por lo tanto, este estudio recopila datos en español relacionados con el coronavirus para proceder después a construir un corpus de coronavirus en español e incrustar palabras a través de un preprocesamiento adecuado. Posteriormente, nos gustaría presentar un corpus e incrustación de palabras adecuadas para coronavirus mediante la comparación de la similitud del coseno y la evaluación de visualización con el corpus español existente.
Palabras Clave: corpus, word embedding, coronavirus.

1 Introduction

Natural Language Processing is the field by which computers understand and utilize human language. Part-of-speech tagging for syntax analysis, parsing, entity name recognition for semantic analysis, document classification, and machine translation are only some of the applications and components necessary for Natural Language Processing. Successful Natural Language Processing requires that the human language be converted in such a way so that it can be understood by a computer. Word embedding, which is the most representative method to do so, is defined as transforming a word into a vector that can be calculated.

Various fields in Natural Language Processing are often established by natural

ISSN 1135-5948. DOI 10.26342/2021-67-3

divisions of human life, developing specialty fields of study according to need. A representative example of this is Natural Language Processing research in the field of biomedical and medical science. Recently, research in biomedical fields have increased exponentially. Accordingly, interest in an efficient approach to biological, chemical, and medical data, such as biomedical papers, patent documents, or electronic medical records are also increasing. However, according to Cohen et al. (2014), in the study of Natural Language Processing in the biomedical field, there are many terms that are not part of the common lexicon, such as certain more obscure or complex ailments, symptoms, medications, and terms that are used with meanings that differ from those used in daily life or those used mainly as abbreviations. Existing corpora or word embedding methods have many limitations due to specific medical terminology, and various studies are being conducted to overcome them.

However, most Spanish corpora and biomedical word embeddings do not accurately reflect the worldwide coronavirus pandemic of 2020. The coronavirus, having only begun in 2019, was quickly transmitted across borders resulting in devastating mortality rates. Currently existing Spanish corpora lack terms related to coronavirus as well as terms related to other infectious diseases such as Middle East Respiratory Syndrome (MERS), which was previously prevalent. Considering the fact that coronavirus is affecting not only biomedical fields but also various social fields around the world, the necessity of constructing a corpus that reflects the terms of epidemiology, including coronavirus, is significant.

Therefore, this study first attempts to compile a coronavirus-related Spanish corpus. This Spanish language coronavirus corpus aims to include data from biomedical and various other related fields. Second, word embedding is implemented based on this corpus, and its performance is verified through comparison and visualization evaluation of cosine similarity with the existing Spanish corpora.

The structure of this paper is as follows: Section 2 describes previous studies relevant to this study, Section 3 details the data and methodology used in this study, Section 4 describes the experiment and its results, and in Section 5 this author presents the conclusions of this study.

2 Previous Studies

2.1 Word Embedding

Word embedding is the process by which text is translated into a representative vector that can be calculated and understood by a computer. This process is widely used in Natural Language Processing in combination with deep learning models by considering not only the meaning of a singular word but also the information that can be understood through a phrase. Word embedding is largely divided into sparse expressions and dense expressions.

Sparse expressions can be represented using either the Bag-of-Word model of expression classification or the TF-IDF model. Bag-of-Word is a method of information retrieval that considers only the frequency of occurrence, whereas TF-IDF considers the frequency of occurrence but then calculates relevance to offset errors that occur due to irrelevant but nevertheless commonly appearing words. This is accomplished by calculating the term frequency (TF) of a text but then incorporating an inverse document frequency (IDF) parameter that assigns low weight to general but high frequency words and high weight to more meaningful terms.

On the other hand, the most representative method of classifying dense expressions from word embedding processes is word2vec (Mikolov et al., 2013). Word2vec is a deep learning-based technique that reconstructs the context of words through dense vector expression. This technique is once again divided into Cbow and Skip-Gram models. Cbow predicts intermediate words using surrounding words, while Skip-Gram predicts surrounding words with intermediate words (Mikolov et al., 2013).

Additionally, there is the FastText extension (Bojanowski et al., 2017), which differs from word2vec in that it assumes that there are various sub-words within a single word. FastText has the advantage of being able to calculate the similarity of new words by its classification of sub-words and then recognizing those sub-words elsewhere. Consequently, even words with low frequency can be embedded efficiently as long as it contains previously classified or recognized sub-words.

2.2 Spanish Corpus and Biomedical Data

Two corpora used in Spanish Natural Language Processing are CORPES XXI[1] and SBWCE[2] (Spanish Billion Word Corpus). Both include both Peninsula/European Spanish and the varieties of Spanish spoken in Latin America.

Additionally, there are various Spanish data sets available in the biomedical field. IBECS[3] (The Spanish Bibliographical Index in Health Sciences) is a dataset built by examining various journals in the health and medical fields. SciELO (Scientific Electronic Library Online)[4] is a dataset of various scientific journals published in Latin America, South Africa, and Spain.

2.3 Spanish Biomedical Word Embedding

Spanish language word embedding is applied in various fields, and research in biomedical fields is also being actively conducted. Segura-Bedmar and Martínez (2017) employed word embedding using the SBWCE corpus, a pre-existing corpus, to simplify drug description data, but there was a limitation in that they did not build a special separate corpus for their resulting data set. Villegas et al. (2018) also collected biomedical data through text mining, but also experienced limitations due to not having obtained word embedding. Meanwhile, Santiso (2018) did obtain word embedding through Spanish medical records. Although this word embedding in biomedical fields was attempted using Electronic Health Records (EHRs) of Spanish hospitals, there is a disadvantage in that no intrinsic evaluation was performed.

According to Sores et al. (2019), in the biomedical field, terms are often used in a manner inconsistent with the typical dictionary definition, or with a different connotation, such as in the case of the word ‘positive’. Or, for example, when two words are combined to represent a single meaning as they are in open compound words, for example: ‘brain dead’. Therefore, in this case, the word2vec method is not suitable for producing an accurate word embedding in the biomedical field because it would recognize “brain” and “dead” separately. To solve this problem, Sores used FastText for biomedical word embedding. FastText is effective in overcoming the above-mentioned limitations as it analyzes the morphological and semantic characteristics of the words by breaking them down and analyzing them in N-grams as opposed to as a whole. This study not only accomplished the feat of obtaining a Spanish word embedding in the biomedical field, but also overcame the prior limitations of word embedding in the Spanish biomedical field by performing both the intrinsic evaluation and the extrinsic evaluation through the recognition of the entity name, which had not been done before.

Following this, Rivera & Martineza (2020) implemented a context-based word embedding and performed entity name recognition using a deep learning neural network model in Spanish medical records. Data was collected from the existing Spanish biomedical corpora and the biomedical corpora of other languages, and word embedding was obtained using the vector contextualizing Bidirectional Encoder Representations from Transformers (BERT) model, not the previously mentioned models word2vec and FastText, which are context-free and therefore may result in less precise word embeddings.

Ever since this advance in technology it has been the objective of this study to overcome the limitations of the currently existing context-free Spanish word embedding models by performing a Named Entity Recognition (NER) using the available word embedded data and deal with the context-based Spanish biomedical word embedding generation process in the future.

In the case of Gutiérrez-Fandiño et.al. (2021), the Spanish biomedical corpus and the Spanish medical record corpus were embedded using the FastText method. Since the former is larger than the latter, embedding was performed separately, and for lower words, embedding was performed using Byte Pair Encoding. Corpus data was

[1] CORPES is a corpus created by Real Academia Española, Spain, and contains more than 300,000 documents with approximately 312 million words derived from written texts and oral manuscripts. It is a corpus that includes literary works such as novels, films, scripts, and plays, as well as words from non-literary books and periodicals, blogs, and Internet resources. Peninsular Spanish data accounts for about 30% of the resources and Latin American Spanish data for about 70% (Corpes XXI”. Real Academia Española. https://www.rae.es/banco-de-datos/corpes-xxi, December 21, 2020)

[2] SBWCE is a corpus created by the University of Chile which contains approximately 1 billion words. It is a Spanish language corpus made from various existing Spanish corpora and is embedded using the word2vec algorithm.

[3] http://ibecs.isciii.es

[4] https://SciELO.org/es/

collected from crawls, books, SciELO, Pubmed, and patent documents. This study is meaningful in that it has created and embedded a corpus large enough to be used in the future in Natural Language Processing in the biomedical field. However, compared to the existing corpus, it is limited in that it has undergone no intrinsic or extrinsic evaluation.

In short, research on word embedding in the field of Spanish biomedical science is indeed progressing, despite the disadvantages of that have shadowed the many advances. However, research in the field of epidemiology, including the all too topical coronavirus, has not been conducted well, and there is a constraint in that, in many cases, the existing biomedical corpus does not contain terms related to infectious diseases. Furthermore, although there are some corpora which reflect words related to coronavirus, an objective evaluation has not been conducted.

3 Data and Method

The objective of this author's study is to construct a corpus and word embedding that effectively reveals information related to infectious diseases, especially in regard to the coronavirus. Therefore, data from various fields related to coronavirus and infectious diseases was gathered, pre-processed, and then a corpus was built, and word embedding was performed through FastText. This study chose the FastText model as its embedding method due to its ability to effectively analyze and embed terms whose biomedical definitions are different from those in use in the common lexicon. In addition, FastText can implement effective embedding even with a small number of words as its data set. The performance of this word embedding was then tested through a cosine similarity evaluation between the existing Spanish corpus and embedded SPWCE. Additionally, the corpus and embeddings will be tested for efficacy of related word construction through visualization evaluation.

3.1 Constructing Corpus and Pre-Processing Data

For the purposes of this study, coronavirus-related data was divided into two categories: bio-medical data and social data. The reason this study included both bio-medical data and social data in the corpus was to create a corpus that reflects various issues which included, not only the coronavirus-related biomedical domain, but also various social, economic, and cultural domains as well. Data in the field of biomedical science was extracted from medical journal articles containing keywords related to coronavirus and other infectious diseases such as "COVID 19", "SARS", or "MERS" published in Spanish-speaking countries. In addition to this collected data, health science data from the IBEC and SciELO data sets was used as well.

Social sector data was obtained from online major daily newspapers from Spain, Mexico, Chile, Peru, Colombia, and Argentina. These articles contained the keywords "Coronavirus", "Covid-19", "SARS", and "MERS". To supplement this data a web-crawler was used to collect data from websites such as Wikipedia.

Following data collection, the text was standardized for processing. Capital letters were changed to lowercase letters, and special letters were removed. Stop words were removed using the NLTK Spanish Stop Word Corpus. In addition, a pre-processing procedure was performed to remove all English words using the NLTK English corpus so that English was not included in the final results. Thus, the Spanish data related to coronavirus was constructed as follows.

	Number of Token
Academia Data	29,057,255
News Paper	124,368,426
Wikipedia	19,374,235
Total	172,799,916

Table 1: Spanish Coronavirus Data Collection Results.

3.2 Word Embedding and Training

FastText is an opensource embedding method. It was developed by Facebook as an alternative way to turn words into vectors. Developed after word2vec, it exhibits similar functions to the skip-gram model of word2vec and the mechanism of CBOW. However, word2Vec sees a word as an indivisible unit, whereas FastText sees that there are smaller sub-words within

words. In FastText, each word can be represented by a set of N-grams made of letters, and after the learning process of the artificial neural network is complete, each N-gram of all the words in the dataset is also embedded (Bojanowski et al., 2017). For example, in the case of "virus", an example of FastText's syllable N-gram function is as follows (where n=3).

(1) $G_{Coronavirus}$ = {<vi, vir, iru, rus, us>, <virus>}

The new score function of FastText is as follows.

$$s(w,c) = \sum_{g \in g_w} z_g^T v_c \qquad (1)$$

In the above equation, z_g^T is the vector of each word in N-gram and v_c is the word vector included in the context.

It is by virtue of this function that it is possible to calculate a word's degree of similarity with other words, even in the case of unlearned words. The FastText Model of word embedding is therefore superior to Word2Vec, at least in regard to the latter's inability to cope with unknown words or to accurately embed words with low frequency within a word set.

3.3 Evaluation

The resulting corpus and word embedding of this study were evaluated through word similarity and visual evaluations.

3.3.1 Word Similarity Test

The word similarity evaluation is a method to evaluate the quality of word embeddings. First, a series of word pairs is created. Then an evaluation of the similarity of the word pairs is conducted by human evaluators. After that the correlation between the scores obtained through these evaluations and the cosine similarity between words and vectors is calculated. This study conducted the following two such evaluations of word similarity.

First, WordSim data created for general word embedding evaluation is selected. WordSim data was human evaluated by dividing the degree of similarity between words by 0-10 points, with a total of 322 evaluation sets. In this study, WordSim data written in English was first translated using Google Translator, and if the meaning of the machine translated data was inaccurate when compared to the original WordSim data, the researcher directly translated the WordSim data set themselves.

Secondly, MayoSRS (Pakhomov et al., 2011) was used to evaluate the word embedding from the biomedical field. MayoSRS is composed of 101-word pairs, and the similarity was also directly human-evaluated with a parameter of 0-10 points.

Third, cosine similarity comparison was performed based on UMSRS-similarity (UMSRS-sim), which was widely used as a similarity comparison set in another biomedical domain (Pakhomov et al., 2010). This data selected consisted of 566-word pairs from a data set in which the similarity of UMLS (Unified Medical Language System) concepts were manually evaluated. These two datasets were also originally written in English, and, like WordSim, after an initial translation through Google Translator, if the resulting word was different compared to the original data, it was directly translated and evaluated using the Spanish version of MayoSRS and UMSRS-sim.

3.3.2 Visualization Test

Following the word similarity evaluation, a visualization evaluation was performed for the internal evaluation of the word embedding. A visualization evaluation is another method of evaluating word embeddings, and it is a technique in which words with similar meanings appear in close proximity to pictures so that humans can easily understand them. Through this method, the quality of the embedding can be checked indirectly for quality. Since word embedding is usually a high-dimensional vector, this study applied t-SNE (t-Stochastic Neighbor Embedding)[5], reduced it to two dimensions, and evaluated the visualization.

[5] t-SNE expresses high dimensional data in a two-dimensional plane by use of an algorithm that preserves the structure of neighboring data and distance as much as possible. In this way researchers may visually analyze their word embeddings.

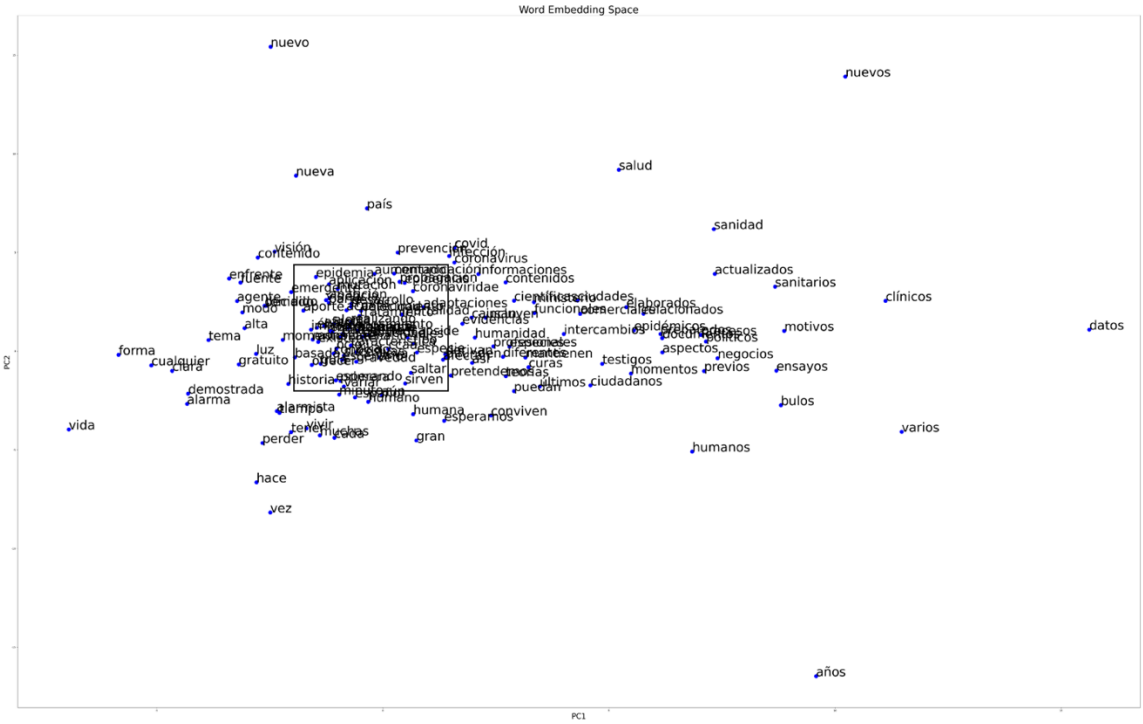

Figure 1: Visualization test result.

4 Result

4.1 Cosine Similarity Test

The word embedding underwent a cosine similarity evaluation with SBWCE, an existing Spanish language corpus. The results of this evaluation were conducted mainly using the English language data sets for cosine similarity evaluation, Word-sim data and MayoSRS, the results of which are shown in <Table 2>.

SCC (Spanish Corona Corpus), the corpus created by this study, showed higher cosine similarity than the existing SBWCE corona. In the case of Word-Sim, which is composed of pairs of common words, the cosine similarity of SCC was 0.4796. This is much higher than that of SBWCE which is 0.2674. In the case of the MayoSRS evaluation, which was composed of medical terms, the cosine similarity of SCC was 0.2025. Compare this again with the cosine similarity of SBWCE which was 0.1174. Once again SCC proves higher. Similarly, in the case of UMSRS-sim composed of biomedical terms, while the similarity of SBWCE was 0.4280, the cosine similarity of SCC was 0.4873, definitively higher in the corpus of this study.

Finally, similarity evaluation was performed using the Multi-SimLex dataset[6]. Although this evaluation resulted in much lower scores than than other similarity evaluation test sets, SCC, the corpus of this study, showed better results than the existing SBWCE.

	SCC (our corpus)	SBWCE
Word-Sim	0.4796	0.2674
MayoSRS	0.2025	0.1174
UMSRS-sim	0.4873	0.4280
multisimlex	0.0981	0.0521

Table 2: Cosine similarity test between SCC and SBWCE.

[6] Multi-SimLex (https://multisimlex.com/) is a data set created to evaluate semantic similarity, and in the case of Spanish, it consists of about 1900 pairs of words. This evaluation was conducted by ten individuals.

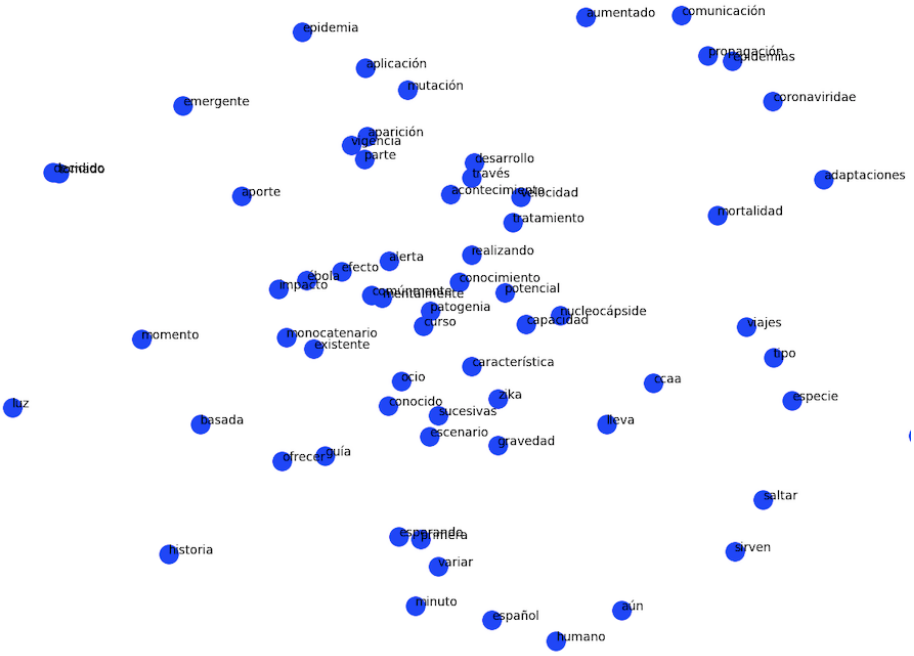

Figure 2: Zoom in overlapping part of Figure 1.

These results conclusively show that the coronavirus data-based corpus conducted in this study demonstrate better performance than the existing Spanish language corpus. Particularly noteworthy is that despite the SCC's fewer embedded words than SBWCE, it has shown a consistently higher performance, strong evidence that the corpus and word embedding in this study are greatly efficient.

4.2 Visualization Test

As a result of the word embedding visualization of the Spanish Corona Corpus in this study (Figure 1 & Figure 2), many epidemiological terms that do not exist in the existing Spanish corpus (SBWCE, etc.) such as "MERS", "SARS", and "Coronavirus" appear. Additionally, various medication terminology related to the treatment of illness in patients appear. For example, terms such as, "epidemiológica", "contagiado", and "positivo" are located close to each other. Words such as, "cuarentena", "crisis", and "economía", which are related to the economic problems of the coronavirus are also concentrated in close proximity. Terms related to social issues in connection to coronavirus are easily displayed. It seems to have been reflected in the results of the visualization.

Additionally, other various words are derived through the visualization evaluation such as "marzo", "brote", "online", and "colegio" (elementary school), etc. In short, is possible to understand the impact of coronavirus on people in disparate areas of society, from economic sectors to education. From the results of the visualization, it can easily be seen that the positions between common similar words are visually near each other.

5 Conclusion

The coronavirus infection (Covid-19) has greatly impacted the world, and the need to process coronavirus-related data has grown ever more urgent. Due to the absence of a dedicated Spanish language coronavirus corpus with an implemented word embedding, this study is essential to the future to the field of Natural Language Processing, specifically in the field of epidemiology and biomedicine, in that it can be used for future related studies by presenting coronavirus-related corpora and word embeddings.

The coronavirus corpus constructed through this study contains various epidemiological terms that are not included in the existing biomedical corpus and word embedding. This dedicated coronavirus corpus has a high cosine similarity between similar words compared to the existing Spanish language corpus making

effective word embedding possible. From the evidence shown here word embedding has been successfully implemented. In addition, according to the results of the visualization evaluation, words on topics related to the fields of medicine, economics, society, education, and miscellany are clustered, providing researchers with the opportunity to understand what the coronavirus specifically suggests in each field. Most importantly, the word embedding created in this study has shown a consistently superior performance alongside the SCC despite its smaller size and serves as proof of its efficacy.

However, the data in this study limited to biomedical academic papers and major daily newspapers sourced from Spain, Mexico, Chile, Peru, and Argentina. It is necessary to secure additional data such as various online journals, blogs, SNS, and books in the future. Furthermore, there is a need for supplementary experiments for more effective word embedding implementation. First, by running several word embeddings using the same model consecutively but redefining the parameters and dimensions of its requirements. Secondly, by obtaining word embeddings using several different models altogether.

References

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135-146.

Cohen, K. B., and D. Demner-Fushman. 2014. *Biomedical Natural Language Processing*, 11. John Benjamins Publishing Company.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, C. P. Carrino, O. De Gibert, A. Gonzalez-Agirre, and M. Villegas. 2021. Spanish Biomedical and Clinical Language Embeddings. arXiv preprint arXiv:2102.12843.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Pakhomov, S., B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. B. Melton. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. *AMIA annual symposium proceedings*, 2010:572.

Pakhomov, S. V., T. Pedersen, B. McInnes, G. B. Melton, A. Ruggieri, and C. G. Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics*, 44(2):251-265.

Rivera-Zavalaa, R., and P. Martineza. 2020. Deep Neural Model with Contextualized-word Embeddings for Named Entity Recognition in Spanish Clinical Text. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

Santiso, S., A. Casillas, A. Pérez, and M. Oronoz. 2019. Word embeddings for negation detection in health records written in Spanish. *Soft Computing*, 23:10969-10975.

Segura-Bedmar, I., and P. Martínez. 2017. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of biomedical semantics*, 8(1):1-9.

Soares, F., M. Villegas, A. Gonzalez-Agirre, M. Krallinger, and J. Armengol-Estapé, 2019. Medical word embeddings for Spanish: Development and evaluation. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 123-133.

Villegas, M., A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, and M. Krallinger. 2018. The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. *LREC MultilingualBIO: Multilingual Biomedical Text Processing*, pages 32-39.