A Systematic Literature Review on Word Embeddings

Luis Gutierrez[1] and Brian Keith[1]

[1] Department of Computing & Systems Engineering
Universidad Católica del Norte
Av. Angamos 0610, Antofagasta, Chile
{luis.gutierrez, brian.keith}@ucn.cl

Abstract. This article presents a systematic literature review on word embeddings within the field of natural language processing and text processing. A search and classification of 140 articles on proposals of word embeddings or their application was carried out from three different sources. Word embeddings have been widely adopted with satisfactory results in natural language processing tasks in general and other domains with good results. In this paper, we report the hegemony of word embeddings based on neural models over those generated by matrix factorization (i.e., variants of word2vec). Finally, despite the good performance of word embeddings, some drawbacks and their respective solution proposals are identified, such as the lack of interpretability of the real values that make up the embedded vectors.

Keywords: Bayesian Networks, Sentiment Analysis, Literature Review, Opinion Mining

1 Introduction

This work is contextualized within the development of a thesis in the field of sentiment analysis, also known as opinion mining. This field focuses on the task of detecting, extracting, and classifying opinions, sentiments, and attitudes concerning different topics in a text. Sentiment analysis has a series of applications, such as the study of political movements in social networks, customer satisfaction, market intelligence, among others [1]. In this context, it is an interest of this systematic review to research vector representations of words to be inserted in semantic vector spaces, with the purpose of providing a formal support for the selection of representations in subsequent works of sentiment analysis. However, it should be noted that the results obtained in this review are also applicable to other fields that lie under the umbrella of natural language processing.

In some studies, the dimensionality of the sparse word-context matrix is reduced through techniques such as singular value decomposition (SVD) [2]. Recently, proposals have been made to represent words through dense vectors that are derived from various training methods, inspired by the modeling of languages through neural networks. These types of representations are referred to as “neural embeddings” or “word embeddings” [2].

Word embeddings have proven to be a mechanism with which the task of computing similarities between words is facilitated by performing efficient calculations through operations with low-dimensional matrices. On the other hand, they are efficient to train, highly scalable for large corpora (thousands of millions of words), and for vocabularies and contexts of similar proportions [2]. The justification for a systematic review lies in the need to summarize or synthesize existing information on a particular topic in an unbiased way [3]and replicable by future interested researchers. In this case, the topic of interest is the application of dense vector representations for words, previously mentioned as word embeddings. The elaboration of a systematic review on this topic becomes relevant, and also necessary, since dense representations of words in semantic vector spaces have played an important role for basic tasks of natural language processing [4, 5], as well as for more complex and pertinent tasks to this work, as is sentiment analysis [6, 7]. On the other hand, no previous systematic reviews were found whose main topic is word embeddings, therefore, this systematic review could be the starting point for subsequent studies.

2 Methodology

The systematic literature review was conducted according to the method proposed in [3], which involves the planning, execution and reporting stages of the research.

2.1 Research planning

Research planning considers the following elements:

1. **Research question**: The fundamental research question of this systematic review is: What semantic vector space representations for words have been proposed and in what contexts?
2. **Keywords**: During the development of the investigation, a set of keywords in English, listed in Table 1 together with an approximate translation into Spanish were considered, since some of them do not have a direct translation. In addition, it is indicated if the plural form of the keyword in English was also considered for the investigation.

Table 1. Keywords and their translation.

English	Spanish	Plural
Word embedding	Incrustamiento de palabra	Yes
Word representation	Representación de palabra	Yes
Vector space	Espacio vectorial	Yes
Semantic	Semántico	No

3. **Selection of sources**: As source selection criteria, the sources used were those whose search engines for articles, accessible within the bibliographic databases of the University, provided scientific articles entirely free of charge, and not only their abstracts and/or bibliographical references. Specifically, the following sources of articles were considered: ACM Digital Library, Scopus, and Web of Science. Most of the articles present in these sources are in English; however, articles in Spanish have also been considered in the application of the selection criteria.
4. **Query string**: A search string was built in order to enter a query into the search engines available in the selected sources of articles, containing all the keywords mentioned above. The search string is as follows: ("word embeddings" OR "word embedding" OR "word representation" OR "word representations") AND ("vector space" OR "vector spaces") AND "semantic".

In addition to searches in the article sources according to the query string, a non-systematic search was also carried out considering bibliographic references of a set of articles, and opportunistic searches on the Internet, as proposed in [8], whose main topic are the representations of words in semantic vector spaces.

2.2 Study selection criteria

Once the initial search has been carried out, the study selection criteria are aimed at identifying those primary studies that provide evidence to answer the research question [3] of this work. Primary studies are understood as studies that contribute to systematic reviews; in contrast, systematic reviews themselves constitute secondary studies. In order to reduce possible biases that may arise, the selection criteria must be defined during the definition of the research protocol [3], and not after planning. The selection criteria of the primary studies are divided into Inclusion Criteria (CI), and Exclusion Criteria (CE). Some criteria, both Inclusion and Exclusion, were based and adapted from [9]. Figure 1 shows the flow chart with the sequential application of the inclusion criteria that were proposed, as well as their interaction with the exclusion criteria.

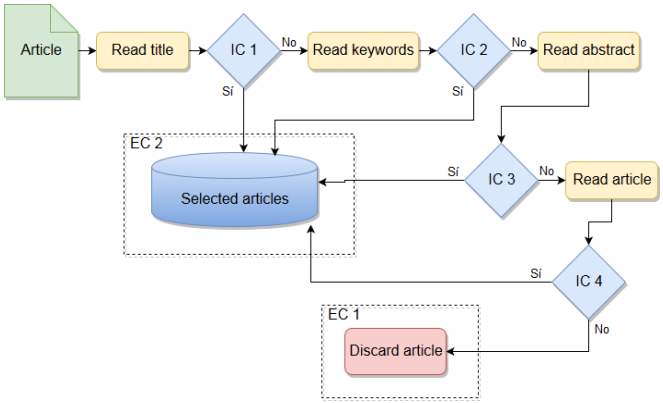

Fig. 1. Flowchart for selecting primary studies.

The Inclusion Criteria (IC) for each article were applied in the sequential order detailed below.

- IC 1: Articles whose title maintains a relationship with some or all the keywords established in this document will be included.
- IC 2: Articles whose keywords are a subset of the keywords established in this document will be included.
- IC 3: Those articles whose abstracts have descriptions or references that deal with representations of words in semantic vector spaces will be included.
- IC 4: Articles that present proposals on new models that generate representations of words in semantic vector spaces will be included.

The Exclusion Criteria (EC) are the following:

- EC 1: All articles that do not comply with any Inclusion Criteria, applied sequentially, will be excluded.
- EC 2: All articles that have already been reviewed in other sources will be excluded.

3 Results and Discussion

3.1 Results of the review

As a first step in the execution of the systematic review, we proceeded to consult the sources of articles selected for the study with the search string constructed and previously presented. The results are shown in the second column of Table 2. The proportion of articles by source, together with the corresponding percentages, is also illustrated in Figure 2-(a).

After reading and applying both the inclusion and exclusion criteria, the articles pertinent to the purpose of this work were identified. The number of selected articles is shown in the third column of Table 2 and the proportion is shown graphically in Figure 2-(b).

Table 2. Number of found and selected articles by source.

Source	Number of found articles	Number of selected articles
ACM Digital Library	14	7
Scopus	125	34
Web of Science	11	4
Total	150	45

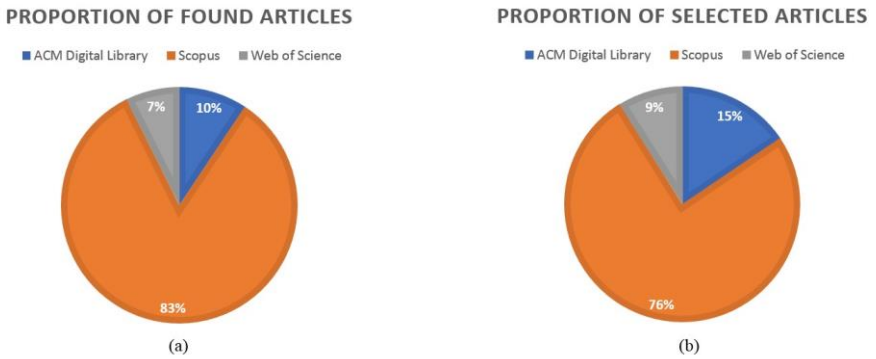

Fig. 2. (a) Graph of number of publications found by source. (b) Graph of number of publications selected by source.

3.2 Discussion

The 45 articles selected in this systematic review position word embeddings as a ubiquitous technique in sentiment analysis, information retrieval, and natural language processing in general. The relevant articles were defined as those that either had a new theoretical proposal to generate word embeddings or those articles that included the concept of word embeddings in disciplines that were not usually directly related with natural language processing and the usage of word embeddings in conjunction with machine learning techniques. The use that the articles made of the concept of word embeddings are mentioned below, beginning with the two representations of words found in almost all the articles, used directly for the realization of the tasks or as a baseline for the evaluation of new proposals of word representations.

word2vec is the name by which are known two language models based on neural networks proposed by Mikolov et al. [10] to generate dense vector representations of words. Specifically, two model architectures are proposed: Continuous Bag of Words Model (CBOW), and Continuous Skip Gram Model (SG). On the one hand, CBOW aims to predict the occurrence of a word given other words that constitute its context. The context of a word w_i is understood as the vicinity composed by the k words to the left of w_i, and the k words to the right of w_i. On the other hand, SG deals with predicting a context given the word w_i. k is a hyperparameter of the model, known as the size of the local context window. In both cases, the models provide dense vector representations for the words that have proved effective in preserving the semantic characteristics of these despite a drastic reduction in dimensionality and training in a shallow neural network, which also speeds up the training process. Following the structure of this document, the works that are related to word2vec are not cited, because as mentioned above, they are almost all, which shows the success of these models.

The second vector representation that appears more recurrently in articles is the so-called Global Vectors for Word Representation (GloVe), proposed by Pennington et al. [11]. Unlike word2vec, which is a predictive model, GloVe is closer to a model that reduces the dimensionality of a co-occurrence matrix of the word-word type,

generated by a fixed dimension local context window. GloVe gets its name from the fact that the statistics of the entire corpus (at a global level) are captured directly by the model. In addition, it proved to be competitive and reported better results than other state-of-the-art methods, such as word2vec, in tasks such as word analogy, word similarity, and Named Entity Recognition.

One of the main uses reported for word embedding is to perform semantic similarity estimation between words of different languages, which in essence goes back to the first applications of natural language processing. In this sense, in [12] a model is proposed to jointly learn bilingual embeddings based only on comparable data constituted by aligned documents that are in two different languages; this model was called Bilingual Word Embeddings Skip Gram (BWESG), which induces a multilingual vector space to embed word representations, queries, and even complete documents. In this same line, Glavaš et al. [13] proposes another method to measure textual semantic similarity between documents written in different languages, characterized by being light on resource usage, and that consists of linearly transferring representations of words from a vector space in a language of origin to a vector space in the target language. The word embeddings used in this work are generated using GloVe and CBOW.

There are also proposals for word embeddings for languages that have very particular alphabets, such as Arabic. Soliman et al. [14] propose a pre-trained set of models of word representations in the Arabic language, in order to provide the community with word embeddings generated from different domains, such as tweets, websites and Wikipedia articles in Arab. The disambiguation of words, a recurring task in the processing of natural language, has also been proposed to solve using word embeddings in Arabic; specifically, Laatar et al. [15] propose this solution in order to develop a dictionary that shows the evolution of the meaning and use of Arabic words, which in turn would help safeguard the Arab cultural heritage. It should be noted that these articles are based on word embeddings generated by the word2vec architectures.

Another common use of word embeddings is reported, which consists of their incorporation into recommender systems. In this sense, Musto et al. [16] present a preliminary investigation for the adoption of word embeddings in which both objects and user profiles are embedded in a vector space, to be used by a content-based recommender system. Boratto et al. [17] state that the use of word embeddings in content-based recommender systems is less effective than other collaborative strategies (for example, the decomposition of singular values), so it proposes another approach, defining a vector space in which the similarity between an object that has not been evaluated by the user and those objects that have been evaluated is measured in terms of linear independence, reaching better results than, for example, SVD. Greenstein et al. [18] assert to be able to convert to words the sequence of objects sought by a user and thus be able to project them in a vector space, in such a way that similarity and analogies between objects can be detected. The word embeddings of these articles are generated according to the word2vec and GloVe models.

The use of word embeddings is also reported in conjunction with other machine learning techniques or linguistic resources. Alsuhaibani et al. [19] establish that the methods that generate vector representations of words based purely on information distributed in a corpus, fail to take advantage of the semantic relational structure that

there is between words in concurrent contexts; These structures are detailed in manually elaborated knowledge bases, such as ontologies and semantic lexicons, where the meaning of words is defined by the various relationships that exist between them. Therefore, the corpus is combined with the knowledge bases to generate word embeddings that, when used, present an improvement in performance in the tasks of measuring similarity and analogy of words, obtaining results that support the hypothesis.

On the other hand, Liu [20] proposes that, in addition to generating vector representations of words having the corpus as a source, internal elements of each word, such as morphemes, should also be taken into consideration. For this, two models are proposed to generate word embeddings: Morpheme on Original view and Morpheme on Context view (MOMC) and Morpheme on Context view (MC), which show higher performance in detecting the similarity of words than the models of the baseline, among which was CBOW. Gallo et al. [21] present a method in which the word embeddings generated by word2vec are encoded in images, to later make use of convolutional neural networks (CNN) and perform text classification on the images. The method reported better classification results when compared to the baseline (doc2vec with SVM).

Wild and Stahl [22] present an implementation of Latent Semantic Analysis and its results when generating word representations in vector spaces. Unlike word2vec and GloVe, this method is based on the factorization of matrices, due to the use of singular value decomposition.

One of the main drawbacks of word embeddings is the decrease in the dimensionality of the problem at the expense of the interpretability of the real values that make up the vector representations, which is commonly known as an opaque model. Given this, solutions have been proposed to make the representations interpretable. Among them, Liu et al. [23] propose techniques for visualizing useful semantic and syntactic analogies in various domains, using as base representations those generated by word2vec and GloVe. Andrews [24] points out that the representations learned by the word embeddings models, despite having small dimensions, still make use of an important amount of storage, so the use of Lloyd's algorithm is proposed since it can compress dense representations by a factor of 10 without further penalizing performance. In addition, an efficient factorization method of computing in GPU is presented to obtain representations with greater interpretability, each dimension is coded with a non-negative value; on the other hand, they are also sparse. The tasks of similarity and analogy of words with the compressed representations were evaluated and it was demonstrated that the aspirations of the work were attainable.

Regarding the relevant articles found in opportunistic searches, Joulin et al. [25] present a method to generate word embeddings several orders of magnitude faster than the models resulting through deep learning. The method is similar to CBOW, replacing the word from the context medium with a classification label, obtaining a performance on par with models that take more time to train in tasks such as sentiment analysis and label prediction. Moody [26] presents a method in which the Skip Gram architecture of word2vec is mixed with the modeling of topics in documents that make use of the Latent Dirichlet Allocation technique. The model is called lda2vec and is capable of generating vector representations of words and

documents in the same vector space. Next, Table 3 shows the most important articles on the investigation, considering proposed models that have been widely adopted in the literature to generate word embeddings; reported improvements to these models; the incorporation of these with other natural language processing techniques (e.g. Latent Dirichlet Allocation); and also, possible solutions to the problem of interpretability that these dense vector representations pose. It must be noted that these criteria answer to the necessity of highlighting articles that are appropriate to future studies, according to the context of the authors' work, and not necessarily to the most frequently cited articles.

Table 3. Number of found and selected articles by source.

Year	Title	Authors
2007	Investigating Unstructured Texts with Latent Semantic Analysis	Fridolin Wild, Christina Stahl
2013	Efficient estimation of representations in vector space	Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
2014	GloVe: Global Vectors for Word Representation	Jeffrey Pennington, Richard Socher, Christopher D. Manning
2016	Bag of Tricks for Efficient Text Classification	Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov
2016	Mixing Dirichlet Topic Models and Word Embeddings to make lda2vec	Christopher Moody
2016	Compressing Word Embeddings	Martin Andrews

4 Conclusions

Dense vector representations of words have been widely adopted with satisfactory results in natural language processing tasks in general. In addition, they have been applied in other domains with good results. It is also worth mentioning one of the limitations of this study, namely that the work has been done on only three databases of scientific publications (ACM, Scopus and Web of Science).

On the other hand, the hegemony of word embeddings based on neural models over those generated by matrix factorization is reported; numerous alternatives have been proposed that are variants of a group of neural models: word2vec. In this context, studies have been conducted on the impact of including other techniques together with word embeddings to perform natural language processing tasks, reporting good results.

As future work, applying the inclusion and exclusion criteria, and the systematic review process used in this work to other scientific publications databases could be a starting point in order to extend the scope of this review. This review could be used as the foundation for further analysis of the literature in the medium and long term, considering the volume of research that is done on word embeddings in different areas.

Finally, despite the good performance of word embeddings, some drawbacks and their respective solution proposals are identified, such as the lack of interpretability of the real values that make up the embedded vectors.

Acknowledgments. Research partially funded by the National Commission of Scientific and Technological Research (CONICYT) and the Ministry of Education of the Government of Chile. Project REDI170607: "Multidimensional Bayesian classifiers for the interpretation of text and video emotions".

References

1. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems 89 (2015) 14–46
2. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Volume 2. (2014) 302–308
3. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele University 33(2004) (2004) 1–26
4. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning, ACM (2008) 160–167
5. Zou, W.Y., Socher, R., Cer, D., Manning, C.D.: Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. (2013) 1393–1398
6. Severyn, A., Moschitti, A.: Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM (2015) 959–962
7. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Volume 1. (2014) 1555–1565
8. Carrizo Moreno, D.: Atributos contextuales influyentes en el proceso de educción de requisitos: una exhaustiva revisión de literatura. Ingeniare. Revista Chilena de ingeniería 23(2) (2015) 208–218
9. Díaz, N.D., Zepeda, V.V.: Ejecución de una Revisión Sistemática sobre Gestión de Calidad para Sistemas Multiagente
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). (2014) 1532–1543
12. Vulíc, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, ACM (2015) 363–372
13. Glavaš, G., Franco-Salvador, M., Ponzetto, S.P., Rosso, P.: A resource-light method for cross-lingual semantic textual similarity. Knowledge-Based Systems (2017)
14. Soliman, A.B., Eissa, K., El-Beltagy, S.R.: AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. Procedia Computer Science 117 (2017) 256–265
15. Laatar, R., Aloulou, C., Bilguith, L.H.: Word sense disambiguation of Arabic language with Word Embeddings as part of the Creation of a Historical Dictionary. In: Language Processing and Knowledge Management proceedings, CEUR-WS.org (2017)

16. Musto, C., Semeraro, G., de Gemmis, M., Lops, P.: Learning word embeddings from Wikipedia for content-based recommender systems. In: European Conference on Information Retrieval, Springer (2016) 729–734
17. Boratto, L., Carta, S., Fenu, G., Saia, R.: Representing Items as Word-Embedding Vectors and Generating Recommendations by Measuring their Linear Independence. In: RecSys Posters. (2016)
18. Greenstein-Messica, A., Rokach, L., Friedman, M.: Session-based recommendations using item embedding. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, ACM (2017) 629–633
19. Alsuhaibani, M., Bollegala, D., Maehara, T., Kawarabayashi, K.i.: Jointly learning word embeddings using a corpus and a knowledge base. PloS one 13(3) (2018) e0193094
20. Liu, J.: Morpheme-Enhanced Spectral Word Embedding. In: Proceedings of the International Conference on Software Engineering and Knowledge Engineering. (2017)
21. Gallo, I., Nawaz, S., Calefati, A.: Semantic Text Encoding for Text Classification Using Convolutional Neural Networks. In: Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on. Volume 5., IEEE (2017) 16–21
22. Wild, F., Stahl, C.: Investigating unstructured texts with latent semantic analysis. In: Advances in Data Analysis. Springer (2007) 383–390
23. Liu, S., Bremer, P.T., Thiagarajan, J.J., Srikumar, V., Wang, B., Livnat, Y., Pascucci, V.: Visual Exploration of Semantic Relationships in Neural Word Embeddings. IEEE transactions on visualization and computer graphics 24(1) (2018) 553–562
24. Andrews, M.: Compressing word embeddings. In: International Conference on Neural Information Processing, Springer (2016) 413–422
25. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
26. Moody, C.E.: Mixing Dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:1605.02019 (2016)