ISSN 1870-4069

Improving Text Representations: A Systematic Literature Review

José Hernández Hernández, Guillermo de Jesús Hoyos Rivera,
Efrén Mezura Montes

Universidad Veracruzana,
Instituto de Investigaciones en Inteligencia Artificial,
Mexico

jclementehdzhdz@gmail.com, {ghoyos,emezura}@uv.mx

Abstract. Natural Language Processing (NLP) is the area charged of designing and developing algorithmic techniques to automatically process text aiming to perform specific tasks related to natural language, its use, and interpretation. These tasks must transform the text into numerical representations so that they can be treated by computational models, such as those of Deep Learning (DL) and Machine Learning (ML). However, we have not found so far a literature review about text representation improvement techniques used in NLP. Most papers are somehow centered on describing Neural Network Language Models (NNLMs). This systematic literature review aims to provide researchers with an overview of the different techniques for improving text representations to identify new areas of opportunity.

Keywords: Text representations, natural language processing, language models, word embeddings.

1 Introduction

NLP comprises the set of algorithmic and mathematical methods that are executed by computers to "understand" human language, also known as natural language [4]. NLP is commonly used to process either, voice or written text, and attempts to solve specific tasks, such as language generation, machine translation, question answering, text classification, sentiment analysis, and part of speech tagging, among others.

One of the main problems of NLP is how to represent text on computers so that they can numerically manipulate it. Thanks to DL, in recent years, researchers looking for new NNLMs which return a set or a list of numerical values, have proposed representations known as embeddings [19]. An embedding is a n-dimensional vector representation containing continuous values representing a word, or a set of words, that are part of a document or a sentence.

Such values contain information about the context where the text is found. This information can be taken from lexical resources, or from making a specific task with a learning approach, such as predicting words or sentences, also known as fine-tuning. Most of the time, this information is said to be distributed along the representation embeddings [20].

This outlook corresponds to the so-called distributional hypothesis [13]. It is important to mention that NNLMs are useful for a set of tasks, and they have shown interesting results, but indeed, they require of high computational power, and the training time is long [16]. This is more important if fine-tuning of a pre-trained model is desired.

Therefore, training could be impractical and could fall into over-fitting. For the above-mentioned, it is necessary to improve embeddings, NNLMs, and pre-trained models. This paper aims to review and summarize novel text representations techniques that emerged after the development of NNLMs, such as Word2Vec (Word to Vector representation), GloVe (Global Vectors for Word Representations), and BERT (Bidirectional Encoder Representations from Transformers).

This review shows different perspectives of the state-of-the-art improving representation techniques and, based on them, try to identify opportunity areas. The rest of the paper is organized as follows: Section 2, presents the related works and background of this systematic literature review. Section 3, briefly describes the used method for this review. Section 4, details, in a narrative way, the results found with a synthesis of the different improving text representations techniques. Finally, Section 5, summarizes the conclusions and future work.

2 Background

This review is a meta-syntheses [22], i. e., a search for new theories, concepts, and key subjects that could provide a view for new approaches, with qualitative information about the analyzed research works. As a part of the background, this section briefly describes some of the main surveys and reviews that operated as motivation to do this systematic literature review.

In [20] a survey is conducted, which includes concepts related to pre-trained models and their embeddings. There, the development of the distributional representations is separated in two generations: the first one includes the Word2Vec and GloVe models, while the second includes recurrent and attention models, such as ELMo (Embeddings from Language Models), BERT, GPT (Generative Pre-trained Transformer), and CoVe (Contextualized Word Vectors).

The survey states that pre-trainig is an advantage that could support language representations, convergence speed of models, and also helps to avoid over-fitting. An empirical survey is presented in [24]. Its goal is to describe and test unsupervised models to represent Twitter text. It includes TF-IDF (Term-Frecuency Inverse-Document-Frecuency), Linear Discriminant Analysis (LDA), Word2Vec, GloVe, BERT, XLNet (Extra Long Network), and ELMo, among others.

The authors evaluate the generated representations by using clustering and found that, for example, BERT, which is improved and has many learning parameters, is not necessarily the best one. Other simpler methods such as TF-IDF could be used instead. Another survey, presented in [**?**], describes different strategies to represent text from the symbolic point of view, to the appearance of the distributed representations learning, such as Word2Vec. This survey could serve as an introduction to the text representation techniques in the DL era.

In [1], a survey on word embeddings is presented. The authors describe distributed representations based on vector space models, statistical language modeling, prediction, and count-based models. Models such as TF-IDF, Word2Vec, GloVe, and statistical methods such as Latent Semantic Analysis (LSA) are included. Finally, the idea of improving the results of NLP tasks by tuning models is presented.

Based on [14], Neural Networks (NNs) that generate embeddings or text representations are treated as language models. In this case, the authors describe models such as Word2Vec and classic recurrent models. A very important finding of this survey is that attention models such as BERT are considered better than other text representations.

Finally, in [2], a survey of NNLMs is introduced, where fifty different models, which include shallow, recurrent, convolutional, and attention models, and their variants, are described. The authors of this survey highlight the computational complexity of NNLM, and they propose to generate new strategies by adding common sense and human intuition to improve text representations.

From this related work review, it can be clearly seen that research is mostly interested in describing NNLMs, and highlighting some aspects of the improvements made. This is why this paper focuses on describing other improvement approaches that can be useful to add to future NNLMs implementations. To the best of the authors' knowledge, there is no systematic literature review about the concepts which are presented in this work.

3 Research Method

The research method implemented in this systematic literature review is that expressed in [22], following the steps explained in the following subsections.

Scoping. Aims at responding the next questions: (1) Which are the main NNLM, embeddings, and text representations? (2) Which are the techniques used to improve the representations?

Searching. A sequence of search terms was conducted to identify relevant work, including: (1) text representations, (2) symbolic text representations, (3) numerical text representations, and (4) improvement of text representations, all of them for NLP. Considering this work as a first step to analyze the state-of-the-art in this topic, the search terms were handled through the Google Scholar engine, and the included main databases were IEEExplore, Springer, ACM, Elsevier, and arXiv, with no restriction to finding works in other sources.

Screening. In this stage, a manual screening of the papers was performed, with special emphasis on the abstract and title of the papers.

Eligibility. An in-depth reading was performed to determine the papers eligibility for inclusion. The following information was extracted: (1) main topics, (2) original text representation or NNLM, (3) improving technique, (4), data type used for experiments (word, sentence, paragraph or document), and (5) publication year or last submission year.

Study quality. Finally, the following is the checklist (based on [15]) that was applied to the papers quality assessment:

Table 1. Paper scores of study quality questions.

Paper	Q1	Q2	Q3	Q4	Q5	Q6	Total score	Year
[6]	1	0	0	1	1	0	3	2014
[3]	1	0	1	1	1	1	5	2014
[25]	1	0	1	1	1	1	5	2014
[5]	1	1	1	1	1	0	5	2015
[26]	1	1	0	1	1	1	5	2016
[18]	1	1	0	1	1	1	5	2017
[17]	1	1	0	1	1	0	4	2018
[9]	1	1	0	1	1	1	5	2018
[21]	1	1	1	1	0	0	4	2019
[10]	1	1	0	0	0	1	3	2019
[11]	1	1	1	1	1	1	6	2019
[23]	1	1	0	1	1	0	4	2020
[12]	1	1	1	1	0	1	5	2020
[7]	1	1	1	1	0	0	4	2021

– Q1: Are the aims clearly stated?
– Q2: Is there a comparison with other methods?
– Q3: Are the used data clearly explained?
– Q4: Is it clear what is the technique used to improve the text representation?
– Q5: Are negative findings present?
– Q6: Is it clear what are the future trends in such an improving technique?

The defined checklist has six questions that can be answered with *yes* (1) or *no* (0). Such a checklist is motivated by the research questions and the findings that can generate new areas of opportunity. Table 1, shows the details of the score obtained, and the corresponding publication year or last submission for each paper. It can be clearly seen that the oldest papers were published in 2014, which coincides with the emergence of NNLMs. Selected papers describe improving implementations, which differ from fine-tuning.

4 Narrative of the Results

In this section, the results of the search performed are presented. Also, implicit answers to the research questions are given by comparing and briefly summarizing the information found. In this way, 14 primary studies are considered and described in this section. The information about these papers is focused on improving text representations, NNLMs output vectors, or word embeddings.

We excluded papers that describe text representations without improving them, but it is essential to mention that NNLMs are a clear advance, and give different views of how to process text, using contextual and non-contextual models such as Word2Vec, GloVe, and BERT. From now on, NNLMs embeddings were used to tackle a considerable number of NLP tasks.

Moreover, in various cases those embeddings were improved using different techniques, such as term weighting, retrofitting, and adding sememes into Word Representation Learning (WRL), or simply adding knowledge information into representation vectors.

These tasks are usually known as fine-tuning, but in this review, we take the concept of fine-tuning related to the re-training of an NNLM to solve a specific task, as well as BERT does. The main difference between fine-tuning and improving representations, is that the latter is part of a post-training, i. e., considering a vector representation from an NNLM, and how an extra method can be included to enrich the contained information.

Another view of the improving methods takes place when external information and knowledge, as lexical resources, are involved in model training. In the following paragraphs, improving techniques are briefly described.

Term weighting. Using the original representation vector from the TF-IDF model, and their own data, the authors of [11, 12] enriched the information by applying an algorithm that modified the TF-IDF, and combined original vectors with a weighting algorithm, respectively.

In [11], an algorithm transforms the original TF-IDF embedding into a matrix of weights. Original TF-IDF designates a weight for each term, and in contrast, it is proposed an algorithm that assigns different weights to a single term, considering the classes in which it appears. On the other hand, the authors in [12] used an optimization algorithm to add compactness and expressiveness to vectors at the sentence level.

Taking as a basis Word2Vec, the algorithm adds information about the frequency of terms and it includes a classifier that determines the weights of each sentence. In both cases, the resulting representation of text was employed for a classification task.

Retrofitting. This technique is a way to update resulting vectors from NNLMs by adding semantic and lexical information. The retrofitting approach is firstly described in [5], where authors use lexical resources such as WordNet, FrameNet, and Paraphrase Database (PPDB) to enrich GloVe and Word2Vec embeddings via label propagation. Such lexical resources can be taken as symbolic or knowledge representations of words and sentences.

In [9], an explicit retrofitting approach is described, which incorporates three elements: (1) word embeddings from an NNLM, (2) lexical resources, (3) a learning model, where the first and second element fit the model. The resulting vector contains the mixed information from former both elements. Such a learning model can be a NN that has the optimization task to minimize the similarity distance while maintaining the original embeddings aspects.Considering the first retrofitting paper, in [21] is described a retrofitting process over vectors from ELMo.

Whereas in [26], with medical terms as the language model, applied retrofitting with the purpose of improving the semantic similarity. Finally, in [7], using knowledge graphs, such as those graphs from the Trans (Translation-Based Model) family, and the retrofitting technique on a re-implemented BERT, the authors made biomedical information extraction.

Sememes. Sememes are included in the WRL using an extension of Word2Vec [18]. Here the authors define a sememe as the minimum semantic unit of word meanings. The technique is based on the extension of Word2Vec and the aggregation of the attention mechanism typical in Transformers.

Instead of words as the required context in the original Word2Vec, authors use the HowNet sememes and senses, then the Word2Vec model is trained from scratch using the original formalization.

Statistical methods. These methods are applied to produce embeddings and an improvement over original representation vectors. In [6], the Canonical Correlation Analysis is used at adding information to LSA word representations, by employing semantic and syntactic relations from other languages, such as French, English and Spanish.

A Clustering-based improving technique is implemented in [3], where the algorithm is applied over words that are represented with semantic spaces such as Hyperspace Analogue to Language (HAL), and Correlated Occurrence Analogue to Lexical Semantic (COALS), among other methods which had not been tested in the representation of words, such as LSA.

Knowledge incorporation. With respect to this technique, in [17] the authors incorporate knowledge information at the training of a Word2Vec model. The resulting vectors were evaluated in two steps: (1) predicting the relatedness of sentence pairs, and (2) sentiment classification, also known as sentiment analysis. Another paper about this technique is presented in [25]. In this case, a technique based on Word2Vec and the aggregation of relations information from WordNet and PPDB is jointly trained.

The first training stage encompasses only the representation of words that appear in a data set of text, i. e., tweets, and the latter stage, includes only the relation information to train a model based on Word2Vec. Both stages are mixed to produce better representations. It is relevant to note that the explicit retrofitting [9] and sememes aggregation [18], can be incorporated in this subsection.

Iteresting improvements. This last subsection is focused on describing two techniques, different from the above, used to improve representations. In [10], a combinatorial Genetic Algorithm is used to select vector representations from two different NNLMs, Word2Vec and GloVe, and the chosen one is assigned to only one word.

Finally, in [23], with the objective to replace the position variable (commonly used in BERT), the authors use a generalization of word embeddings through continuous functions. As can be seen, most of the improving techniques use Word2Vec or GloVe as the main NNLMs. On the other hand, WordNet is the knowledge lexical resource that is used to enrich word representations.

Some of the reviewed papers are not clear about what is their improving technique future trend, while the rest propose to continue scaling and improving their technique. These improving efforts could be applied to transformer-based models such as BERT, or even be part of few-shot or zero-shot learning to enrich the input to the models used. Knowledge or symbolic information could be added as a part of a specified task fine-tuning.

5 Conclusions and Future Work

This systematic literature review briefly summarized with a narrative and meta-syntheses way, improving techniques over NNLMs, embeddings, or text representations. Different techniques were found, such as term weighting, retrofitting, knowledge incorporation such as sememes, and two interesting approaches that could provide improvements to NNLMs.

The analyzed research works had the expected level of relevancy, and they follow the main motivation of this review: known improving techniques in text representations. As part of future work, and emphasizing that this study is the first one of its kind, we need to search in-depth new concepts such as WRL, sememes, zero-shot and few-shot learning, because it is possible to find other improving techniques of representations in general, as they can be useful for text representations.

Acknowledgments. The first author acknowledges CONACyT's support to pursue graduate studies.

References

1. Almeida, F., Xexeo, G.: Word Embeddings: A Survey (2019). DOI: 10.48550/ARXIV.1901.09069.
2. Babic, K., Martinčić-Ipšić S., Meštrović , A.: Survey of Neural Text Representation Models. ´ Information, vol. 11, no. 11, pp. 511 (2020). DOI: 10.3390/info11110511.
3. Brychcín, T., Konopík, M.: Semantic Spaces for Improving Language Modeling. Computer Speech and Language, vol. 28, no. 1, pp. 192–209 (2014). DOI: 10.1016/j.csl.2013.05.001.
4. Eisenstein, J.: Introduction Natural Language Processing. The Massachusetts Institute of Technology (2019)
5. Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., Smith, N. A.: Retrofitting Word Vectors to Semantic Lexicons. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1606–1615 (2015). DOI: 10.3115/v1/n15-1184.
6. Faruqui, M., Dyer, C.: Improving Vector Space Word Representations using Multilingual Correlation. In: 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 462–471 (2014). DOI: 10.3115/v1/e14-1049.
7. Fei, H., Ren, Y., Zhang, Y., Ji, D., Liang, X.: Enriching Contextualized language Model from Knowledge Graph for Biomedical Information Extraction. Briefings in Bioinformatics, vol. 22, no. 3, pp. 1–14 (2021). DOI: 10.1093/bib/bbaa110.
8. Ferrone, L., Zanzotto, F. M.: Symbolic, Distributed, and Distributional Representations for Natural Language Processing in the Era of Deep Learning: A Survey. Frontiers in Robotics and Artificial Inteligence, vol. 6 (2020). DOI: 10.3389/frobt.2019.00153.
9. Glavaš, G., Vulić, I.: Explicit Retrofitting of Distributional Word Vectors. In: 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 34–45 (2018). DOI: 10.18653/v1/p18-1004.
10. Gunasegaran, T., Cheah, Y. N.: Evolutionary Combinatorial Optimization for Word Embedding in Sentiment Classification. Malaysian Journal of Computer Science, pp. 34–45 (2019). DOI: 10.22452/mjcs.sp2019no3.3.

11. Guo, B., Zhang, C., Liu, J., Ma, X.: Improving Text Classification with Weighted Word Embeddings Via a Multi-channel TextCNN model. Neurocomputing, vol. 363, pp. 366–374 (2019). DOI: 10.1016/j.neucom.2019.07.052.
12. Gupta, S., Kanchinadam, T., Conathan, D., Fung, G.: Task-Optimized Word Embeddings for Text Classification Representations. Frontiers in Applied Mathematics and Statistics, vol. 5, pp. 1–10 (2020). DOI: 10.3389/fams.2019.00067.
13. Harris, Z. S.: Distributional Structure. WORD, vol. 10, no. 2–3, pp. 146–162 (1954). DOI: 10.1080/00437956.1954.11659520.
14. Jing, K., Xu, J.: A Survey on Neural Network Language Models (2019). DOI: 10.48550/ARXIV. 1906.03591.
15. Kitchenham, B., Charters, S.: Guidelines for Performing Systematic Literature Reviews in Software Engineering (2007)
16. Lasse F. W. A., Kanding, B., Selvan, R.: Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models (2020). DOI: 10.48550/ARXIV.2007.03051.
17. Li, Y., Wei, B., Liu, Y., Yao, L., Chen, H., Yu, J., Zhu, W.: Incorporating Knowledge into Neural Network for Text Representation. Expert Systems with Applications, vol. 96, pp. 103–114 (2018). DOI: 10.1016/j.eswa.2017.11.037.
18. Niu, Y., Xie, R., Liu, Z., Sun, M.: Improved Word Representation Learning with Sememes. In: 55th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 2049–2058 (2017). DOI: 10.18653/v1/P17-1187.
19. Pilehvar, M. T., Collados, J. C.: Embeddings in Natural Language Processing, Springer International Publishing (2021). DOI: 10.1007/978-3-031-02177-0.
20. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained Models for Natural Language Processing: A Survey. Science China Technological Sciences, vol. 63, no. 10, pp. 1872–1897 (2020). DOI: 10.1007/s11431-020-1647-3.
21. Shi, W., Chen, M., Zhou, P., Chang, K. W.: Retrofitting Contextualized Word Embeddings with Paraphrases. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 1198–1203 (2019). DOI: 10.18653/v1/D19-1113.
22. Siddaway, A. P., Wood, A. M., Hedges, L. V.: How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-analyses, and Meta-syntheses. Annual Review of Psychology, vol. 70, pp. 747–770 (2019). DOI: 10.1146/annurev-psych-010418-102803.
23. Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., Simonsen, J. G.: Encoding Word Order in Complex Embeddings (2019). DOI: 10.48550/arXiv.1912.12333.
24. Wang, L., Gao, C., Wei, J., Ma, W., Liu, R., Vosoughi, S.: An Empirical Survey of Unsupervised Text Representation Methods on Twitter Data (2020). DOI: 10.48550/ARXIV.2012. 03468.
25. Yu, M., Dredze, M.: Improving Lexical Embeddings with Semantic Knowledge. In: 52nd Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 545–550 (2014). DOI: 10.3115/v1/p14-2089.
26. Yu, Z., Cohen, T., Bernstam, E. V., Johnson, T. R., Wallace, B. C.: Retrofitting Word Vectors of MeSH Terms to Improve Semantic Similarity Measures. In: 7th International Workshop on Health Text Mining and Information Analysis, pp. 43–51 (2016). DOI: 10.18653/v1/w16-6106.