Experimental Comparison and Scientometric Inspection of Research for Word Embeddings

Minni Jain, Sidharth Bansal, and Yatin Gupta

Abstract Word embedding or universal embeddings are the representation of textual data into vectors of real numbers. It acts as a link between the human understanding of a text to that of the machine. It maps the high-dimensional textual data into a vector space of low dimension, which represents the complex relationships existing in the data. Thus, it boosts the performance of tasks involving natural language processing. To know about the popularity gain and the impact of this field, research literature between 2000 and 2019 is analyzed with the help of a scientometric mapping for research done in word embeddings. The paper visualizes year-wise analysis, demographic analysis, category-wise distribution, and document type-wise distribution of the publications indexed in Web of Science (WoS). The paper also comprises a comparative study of Word2Vec, FastText, global vector representation of words, and bidirectional encoder representations for transformers. Pre-trained models are used for experimental comparison. Performance is measured by calculating the deviation of the similarity score of two words given by the models from manually assigned similarity scores by experts, repeated over a list of words on various datasets. The least deviation is shown by FastText due to the usage of morphological information in the skip-gram model and n-gram architecture.

Keywords Scientometric analysis · Word embeddings · Word2Vec · BERT · GloVe · FastText · WordNet · Web of science

1 Introduction

In the contemporary world, the data is abundant and heterogeneous. There is a need to process and interpret the results from this immense data. For example, the reviews from an e-commerce Web site can be used to predict the relevant products which the

M. Jain · S. Bansal (✉) · Y. Gupta
Department of Computer Science and Engineering, Delhi Technological University, New Delhi, Delhi, India
e-mail: bansal.sidharth2996@gmail.com

A. Abraham et al. (eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks*, Lecture Notes in Networks and Systems 167,
https://doi.org/10.1007/978-981-15-9712-1_1

user wants to buy, understanding the semantics associated with the articles, classifying emails as promotional emails, priority emails, spam emails, social emails, etc. All of these tasks require an understanding of the natural language [1]. As the data is huge, humans find it difficult to interpret results from all the documents. Machines do not understand the words. The computer just understands the binary language or mathematical language. So, machines require special strategies for understanding and analyzing the natural language to derive conclusions [2]. One way to represent words is vectors. These vectors are called embeddings. Thus, word embeddings [3] are a mathematical way to represent natural language words and phrases so that computers can understand the natural language text.

The central inspiration of our analysis is to comprehend the research done in word embeddings, by performing a scientometric analysis and a comparison of different state of the art word embedding models. The scientometric analysis demonstrates the growing use of word embeddings, domains they are primarily used in, countries where they are widely used, etc. Then, different word embedding models are compared on various properties, and an experimental study is performed to compare and quantify their performance. Various word embedding models are proposed to date. For example, Word2Vec [4], one hot encoded vector [5], bidirectional encoder representations from transformers (BERT) [6], global vectors for word representation (GloVe) [7], FastText [8], etc. Word embeddings gained immense popularity due to its usage in machine learning applications. The comparative study enables us to choose the model for these applications.

We used the research publications indexed on the Web of Science [9] for the year 2001 to the year 2019. The paper tried to calculate the interestingness and growth in this field. We used document-type distribution, the demography-based analysis, and category-wise distribution to demonstrate the interestingness. Paper visualizes the rapid growth in research and usage of word embeddings in the last few years. It also shows the domain distribution where word embeddings are used with artificial intelligence leading by a big factor. The widespread demographic spread of these architectures was identified by interpretation of the country-wise distribution. Section 2 discusses a scientometric analysis [10] in the field of word embeddings, and Sects. 3 and 4 compare different word embedding models proposed to date based on properties by using different datasets. The study helps us to demonstrate detailed insights on different word embedding models on various datasets. Section 5 concludes the study and tells areas for future work.

2 Scientometric Analysis

This section comprises the scientometric analysis and mapping done on research papers. It comprises various tables and figures consisting of details of resultant values found in the field of word embeddings from the Web of Science portal.

Table 1. Details of dataset

Source/index	Document types	Category	Years	Total number of papers retrieved	Date of download
Web of science	The article, proceeding paper, early access, review and letter	Artificial intelligence, information systems, software engineering, interdisciplinary application, computer engineering theory methods, computer science hardware architecture, cybernetics	2001–2019	741	25.11.2019

2.1 Details of the Dataset

Table 1 shows the details of the research papers we used for the scientometric analysis. We analyzed a total of 741 research publications for "word embeddings" from 2001 to 2019. Out of which, 740 papers were in the English language, while a single paper was in Spanish.

2.2 Document Type-Wise Distribution

Figure 1 demonstrates a treemap consisting of the distribution of document types.

Fig. 1 Treemap demonstrating document types with their record counts

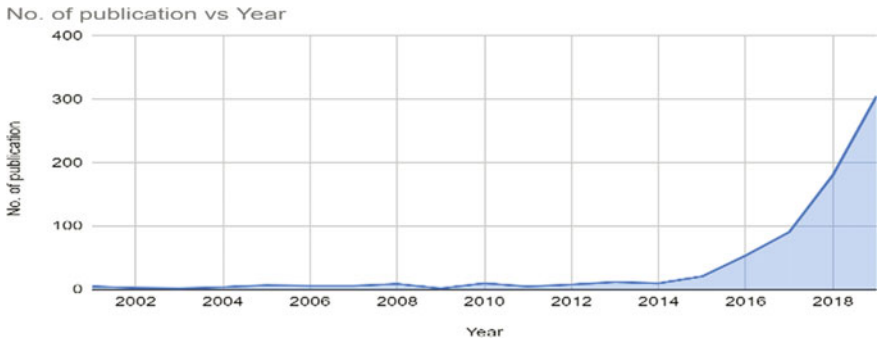

Fig. 2 Graphical representation of the number of papers published per year

Articles are the most commonly used document type. 731 articles are produced until 2019.

2.3 Year-Wise Publications of Word Embeddings

Figure 2 represents an increased number of publications in the field of word embeddings from 2001. The popularity of word embeddings rose to many folds in the past 20 years.

2.4 Demographic Distribution of Word Embedding Research Publications

We analyzed different research publications in the field of word embedding in the year 2001–2019 in Fig. 3. The top three countries contributing to this field were China, the USA, and England. China contributed 292 record counts which is 39.4% of 741 total publications.

2.5 Category-Wise Distribution of Word Embedding Research Papers

The most number of research publications was under "Computer Science Artificial Intelligence" comprising 21.2% of the total distribution. The number of papers was 327. Figure 4 demonstrates these statistics in the form of a pie chart.

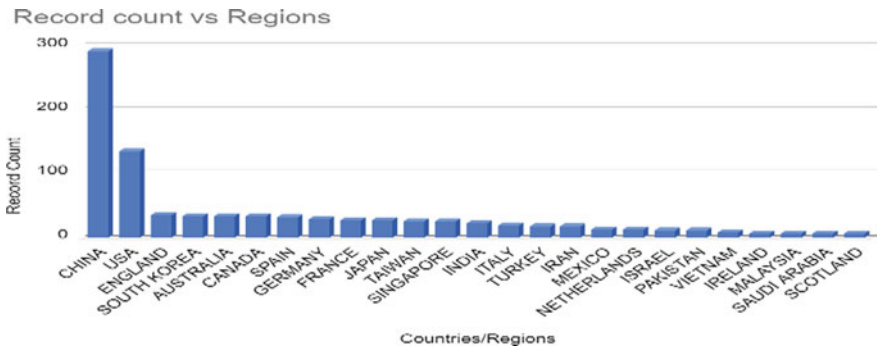

Fig. 3 Country-wise distribution of the publications

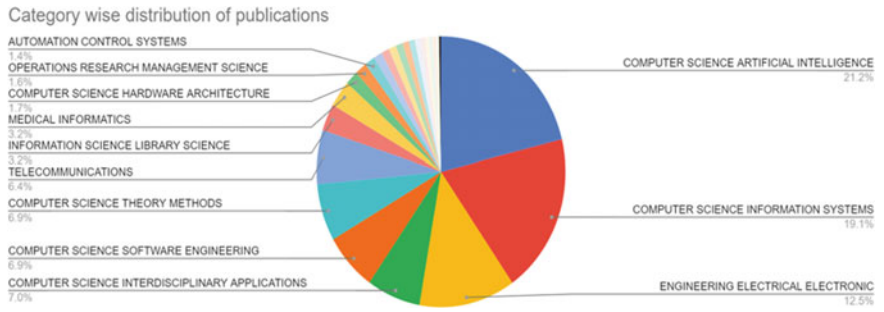

Fig. 4 Category-wise distribution of the publications

3 Comparison of Universal Embeddings Models

Different aspect-based comparison between Word2Vec, global vectors for word representation (GloVe), FastText, and BERT is done in Table 2. The table differentiates between word embedding models based on the technology used, types of the model, applications, advantages, etc.

4 Experimental Study of Word Embeddings Models

Table 3 shows the performance of different word embeddings models on various datasets. Performance is measured by calculating the deviation of the similarity score given by the model for two words from a manually assigned similarity score by experts, repeated over a list of words in different datasets. The lesser value of deviation implies the model similarity scores differ from manually assigned scores very less, and hence, the model is better. Following pre-trained universal embedding models were used. They were trained on the dataset as described below:

Table 2 Difference between word embeddings

S. No.	Property	Word2Vec	GloVe	FastText	BERT
1	Technology used	Neural networks [11]	Matrix factorization techniques [12]	N-grams [13]	Transformer [14]
2	Type of model	Predictive models	Count-based models	Predictive model	Deep learning model
3	Types	Skip-gram and continuous bag of words (CBOW)	GloVe 50D, GloVe 100D, GloVe 200D, and GloVe 300D	FastText with subword info and without subword info	None
4	Application	Question answering, named entity resolution [15], automatic summarization, sentiment analysis [16], etc.	Named entity recognition (NER) tasks, word similarity [17], and word analogy	Content tagging, content classification, sentiment analysis, spam filtering [18]	Next sentence classification [19], named entity recognition (NER), question and answer system
5	Advantages	Word2Vec CBOW tends to produce vectors that are more topically related, skip-gram pays more attention to words in the close proximity and tends to have more syntactic information as a result	Concurrent queries can be processed GloVe tends to produce vectors that are more topically related too	While learning word representations, FastText considers the internal structure of words that are useful for words that occur rarely and morphologically rich languages. Thus, it increases performance	The bidirectional context of the BERT is applied in the reconstruction process
6	Out of vocabulary words and rare words	Word2Vec is unable to represent words that are absent in the training dataset	Cannot handle it	FastText can create words that are absent in the training corpora by using its n-grams	Cannot handle it

- Word2Vec: GoogleNews Word2Vec [20] is used which was trained on 100 billion words and phrases. The length of the vector is 300 dimensions.
- GloVe: This model was trained on Wikipedia 2014 + Gigaword 5 [21]. Gigaword 5 consists of 400 K vocabulary of words and phrases, uncased, 50 dimensional, 100 dimensional, 200 dimensional, and 300-dimensional vectors, 6 B tokens. The dataset is of 822 MB.

Table 3 Deviations of different word embedding models

Dataset	GoogleNews Word2Vec (%)	GloVe 50D (%)	GloVe 100D (%)	GloVe 200D (%)	GloVe 300D (%)	FastText (%)	FastText (with subword info) (%)	BERT (%)
WS-353 [24]	53.514	27.746	29.475	36.251	43.251	27.640	26.883	34.333
WS-353-REL [25]	59.786	29.052	29.641	36.108	44.015	30.837	30.431	44.521
Mc-35 [26]	48.18	55.357	50.894	51.661	55.701	46.732	42.107	71.354
RG 65 [27]	47.633	49.190	45.950	50.846	56.016	41.670	37.548	74.118
Card-660: Cambridge rare word dataset [28]	14.443	14.351	15.877	17.022	18.221	19.255	20.113	59.553
Stanford rare word (RW) similarity dataset [29]	54.189	55.669	59.981	63.741	66.457	41.154	34.026	32.690
MEN [30]	43.769	31.501	32.423	39.075	45.795	29.006	27.668	56.114
MTURK-771 [31]	60.384	31.862	37.100	46.203	54.458	28.420	25.650	36.482

- FastText: Pre-trained on statmt.org news dataset[1] [22] and UMBC corpus. UMBC is a Web-based dataset. statmt.org dataset consists of 16B tokens.
- BERT: Pre-trained on Wikipedia and BookCorpus dataset [23]. It was trained for 1 million epochs.

The FastText model with subword information performs better than the other models due to its ability to understand directly use the morphological information. For instance, word1 = "animal" and word2 = "animals" have the same prefix and similar meanings. However, the words "man" and "management" have different meanings. The relation between the words "animal" and "animals" is the same as the relation between "reptile" and "reptiles". FastText uses this morphological information in the skip-gram model, whereas other models fail to do so. Hence, the deviation of FastText with subword information is lesser than other models.

[1] Facebook Open Source [22].

5 Conclusion

This work discussed the scientometric analysis in the field of word embeddings and a comparison between the various word embedding models. The scientometric analysis showed a tremendous increase in interest in this field over the years. Seven hundred and forty-one publications were found for "word embeddings" between the years 2000–2019 indexed on the Web of Science (WoS). Analyzing the demographic distribution of these publications, China was found the most significant contributor. Also, the topic gained fascination in the USA, England, South Korea, and many other countries. Category-wise distribution showed "Computer Science Artificial Intelligence" having the most publications. The comparative study of universal embedding models showed the different technologies used, advantages, applications, etc., provided by the different models. The experimental comparison using the similarity score generated by different models showed FastText with subword information outperforms other models due to an understanding of the morphological information and n-gram architecture. For future work, machine learning techniques trained on different word embeddings can be applied to distinct problems. Thus, a comparative study to know which word embedding outperforms in which application scenarios can be proposed. This can also be extended to various domains where natural language processing is used like artificial intelligence, virtual reality, information systems, etc. Effect of hyperparameter tuning, differences in semantic and spatial relationships among the words in each of these models can be studied.

References

1. Pauw S, Hilferty J (2016) Embodied cognitive semantics for quantification. Belgian J Linguist 30(1):251–264
2. Lai S, Liu K, He S, Zhao J (2016) How to generate a good word embedding. IEEE Intell Syst 31(6):5–14
3. Word Embeddings (2019) Available from: https://en.wikipedia.org/wiki/Word_embedding. Accessed 25 Nov 2019
4. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
5. Hinton GE (1986, Aug). Learning distributed representations of concepts. In: Proceedings of the eighth annual conference of the cognitive science society, vol 1, p 12
6. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
7. Pennington J, Socher R, Manning CD (2014, Oct) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
8. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. arXiv:1607.01759
9. Available from https://clarivate.com/webofsciencegroup/solutions/web-of-science/. Accessed 26 Nov 2019
10. Piryani R, Madhavi D, Singh VK (2017) Analytical mapping of opinion mining and sentiment analysis research from 2000–2015. Inf Process Manage 53(1):122–150

11. Beale HD, Demuth HB, Hagan MT (1996) Neural network design. Pws, Boston
12. Salle A, Idiart M, Villavicencio A (2016) Matrix factorization using window sampling and negative sampling for improved word representations. arXiv:1606.00819
13. Cavnar WB, Trenkle JM (1994, Apr) N-gram-based text categorization. In: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, vol 161175
14. Dehghani M, Gouws S, Vinyals O, Uszkoreit J, Kaiser Ł (2018) Universal transformers. arXiv:1807.03819
15. Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1):3–26
16. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011, June) Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1. Association for Computational Linguistics, pp 142–150
17. Lin D (1998, July) An information-theoretic definition of similarity. ICML 98:296–304
18. Suchomel V (2017) Removing spam from web corpora through supervised learning using FastText
19. Hassan A, Mahmood A (2018) The convolutional recurrent deep learning model for sentence classification. IEEE Access 6:13949–13957
20. Pre-trained vectors trained on part of Google News dataset (about 100 billion words). Available from https://code.google.com/archive/p/word2vec/. Accessed 27 Nov 2019
21. Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation. Available from https://nlp.stanford.edu/projects/glove/. Accessed 27 Nov 2019
22. Facebook Open Source: FastText: Library for efficient text classification and representation learning. Available from https://fasttext.cc/docs/en/english-vectors.html. Accessed 28 Nov 2019
23. BookCorpus Dataset. Available from https://github.com/sgraaf/Replicate-Toronto-BookCorpus. Accessed 29 Nov 2019
24. The WordSimilarity-353 test collection. Available from https://aclweb.org/aclwiki/WordSimilarity-353_Test_Collection_(State_of_the_art). Accessed 28 Nov 2019
25. WordSim353: Similarity and relatedness. Available from https://alfonseca.org/eng/research/wordsim353.html. Accessed 28 Nov 2019
26. MC-35 Dataset. Available from https://web.eecs.umich.edu/~mihalcea/downloads.html. Accessed 29 Nov 2019
27. RG-65 test collection (state of the art). Available from https://aclweb.org/aclwiki/RG-65_Test_Collection(State_of_the_art). Accessed 29 Nov 2019
28. Pilehvar MT, Kartsaklis D, Prokhorov V, Collier N (2018) Card-660: Cambridge rare word dataset a reliable benchmark for infrequent word representation models. arXiv:1808.09308
29. Stanford rare word (RW) similarity dataset. Available from https://nlp.stanford.edu/~lmthang/morphoNLM/. Accessed 29 Nov 2019
30. The MEN test collection. Available from https://staff.fnwi.uva.nl/e.bruni/MEN. Accessed 30 Nov 2019
31. The word relatedness Mturk-771 test collection. Available from https://www2.mta.ac.il/~gideon/mturk771.html. Accessed 30 Nov 2019