International Journal of Information Technology and Language Studies (IJITLS)

Word Embeddings Based on Spectral Analysis: A Novel Approach

Mohamed Yehia Dahab[1], **Omar A. Batar**[2], **Muazzam Siddiqui**[3], and **Reda Mohamed Salama Khalifa**[2]
mdahab@kau.edu.sa; obatar@kau.edu.sa; maasiddiqui@kau.edu.sa; rkhalifa@kau.edu.sa

[1] Department of Computer Science, Faculty of Computing and Information Technology King Abdul-Aziz University, Jeddah, Saudi Arabia
[2] Department of Information Technology, Faculty of Computing and Information Technology, King Abdul-Aziz University, Jeddah, Saudi Arabia
[3] Department of Information Systems, Faculty of Computing and Information Technology King Abdul-Aziz University, Jeddah, Saudi Arabia

Abstract. Recently, deep learning algorithms have gained huge attention. However, such algorithms are not the optimal solution for many tasks. Spectral analysis transformation algorithms, such as wavelet-transform and Fourier transform, have been successfully applied on many NLP tasks. The challenging issue of using spectral analysis is how to construct a meaningful signal from a text. In word2vec models, different types of neural networks have been applied to learn vector representations of words, which carry the semantic similarities of each word in a specific dataset. Training the word embeddings is computationally very expensive and constrained by the available resources. However, this paper provides an optimized computational complexity for developing word embeddings using parallel computing as well as utilizing an upper ontology to represent the main components of the word vector. Moreover, this research shows how to represent a term as a vector to facilitate computing the similarity or relatedness among different terms. Therefore, this research considers the spectral analysis, which also includes the spatial information of the words around the current word.

Keywords: *word embeddings, spectral analysis, wavelet-transform, term signal, parallel computing*

1. Introduction

Spectral analysis transformation algorithms, such as wavelet-transform and Fourier transform, have been successfully applied for many NLP tasks. Such tasks include information retrieval (Park et al., 2005a; Park et al., 2005b; Dahab et al., 2018a; Dahab et al., 2018b; Dahab et al., 2016; Alnofaie et al., 2016; Aljaloud. et al., 2016; Costa & Melucci, 2010), text classification (Diwali et al., 2015), text clustering (Al-Mofareji et al., 2017), etc. These approaches can also be utilized in developing word embeddings as well.

In word2vec models, different types of neural networks have been applied to learn vector representations of words, which carry the semantic similarities of each word in a specific dataset. Training the word embeddings is computationally very expensive. However, this paper provides an optimized computational complexity for developing word embeddings (Mikolov et al., 2013b; Rong, 2014; Yao et al., 2017).

The state-of-the-art algorithms predict the current word based on the context only, which is the surrounding words given the current word. Therefore, words outside the determined context are disregarded. This research takes into consideration the words outside the determined context but with a lower weight. Moreover, this research considers the spectral analysis of the words around the current word (focal word) rather than the spatial information of words around the current word.

The main aim of this manuscript can be summarized as follows:

- To represent the terms as vectors where semantically related words are mapped to nearby points. This facilitates the application of mathematical operators such as the following famous example (Mikolov et al., 2013a; Drozd et al., 2016; Chen et al., 2016):

$$\boldsymbol{king - man + woman = queen}$$

- To solve the time complexity problem related to the word embeddings, parallel computing has been used as well as applying spectral analysis transformation. To the best knowledge of the authors, this is the first paper that addresses the word embeddings employing the spectral analysis transformation.

- To consider the terms outside the current context with a different weight depending on the distance.

- To determine the dimensionality of the word vector by employing a representative parametric light version - the Suggested Upper Merged Ontology (SUMO) (Niles & Pease, 2001).

The remaining part of the manuscript is outlined as follows: Section (2) presents the related background, Section (3) shows the preprocessing tasks and introduces the proposed work, Section (4) demonstrates the experiments and the datasets used in this research, and finally, the conclusion is given in Section (5).

2. Background and Related Work

In word2vec models, different types of neural networks have been applied to learn vector representations of words, which carry the semantic similarities of each word in a specific dataset. Training the word embeddings is computationally very expensive (Mikolov et al., 2013; Rong, 2014; Yao et al., 2017).

A word vector contains oat values and dimensions, where each value captures a dimension of the word's meaning (Senel et al., 2018). Each dimension represents a meaning, and the numerical value holds on that dimension represents the distance to that meaning. The words that have similar values should have similar or related meanings. In other words, the semantics of a word are embedded through the dimensions of the vector (Levy & Goldberg, 2014). The more the dimensionality of the word vector and the training data, the more accuracy of the extracted semantic relationship (Mikolov et al., 2013a). Many research provided efforts to apply mathematical operators on word vector, such as the following famous example (Mikolov et al., 2013a; Drozd et al., 2016; Chen et al., 2016):

$$\boldsymbol{king - man + woman = queen}$$

Many researches convert text to numerical signal (Park et al., 2005a; Park et al., 2005b; Dahab et al., 2018a; Dahab et al., 2018b; Dahab et al., 2016; Alnofaie et al., 2016; Aljaloud. et al., 2016; Costa & Melucci, 2010; Diwali et al., 2015; Al-Mofareji et al., 2017) and apply signal processing and transformation such as wavelet-transform and Fourier transform in many text mining problems such as text classification, text clustering, and information retrieval. The main idea is how to measure the distances among different terms in a text document, taking into consideration the different occurrence of terms or how to measure the relatedness of a text document to a given query.

3. The Proposed Work

The proximity method calculates the relatedness score based on the distance between two given terms of a specific domain or a text corpus. The larger the distance, the lower the relatedness score and vice versa. The proximity, between two terms t_1 and t_2, can be considered as a measure of terms' dependency (Cummins & O'Riordan, 2009).

The proximity method, that is based on spectral analysis, depends on the spectral domain rather than the spatial domain. To perform this task, after preprocessing the corpora under consideration, a concordance signal is constructed for each two terms, t_1 and t_2, in all documents. Then, the concordance signal is converted into a concordance spectrum using one of the wavelet-transform methods.

3.1. Preprocessing

The preprocessing task includes:

1. Verifying the Unicode used in the text.
2. Tokenizing and assembling all terms in the corpora under consideration into a collection.
3. Stemming each term in the stored collection using the Porter stemming algorithm (Porter, 2008).

4. Calculating the weight of each term using Inverse Document Frequency (IDF).
5. Removing the terms with low weight i.e., stopwords.

3.2. Determining a Representative Parametric Light Version from SUMO

The general part of SUMO contains 1,116 terms, 692 of them are classes (Merge.kif, version 1.75) (Reiter, 2007). SUMO is general enough to cover a wide range of different domains. To determine the dimensionality of the word vector by employing a representative parametric light version of SUMO. Determining the dimensionality of the word vector to form Main Component List (MCL) is based on two conditions:

1. Selecting a subset from SUMO terms that covers all domains.
2. The terms selected should have a high frequency in the corpus under consideration.

Based on past researches, it is preferable for MCL to have from 300 to 800 components.

3.3. Constructing Term Signal

Each document is divided into a predefined number of segments/bins, B. To facilitate computing wavelet-transform, $B = 2^z$ where $z \in N$. If B is very small, this means that many irrelevant words will be found in the same bin and accuracy becomes very low, and vice versa. Also, If B is very large, this indicates that a lot of computations will be needed. Many researches (Park et al., 2005a; Park et al., 2005b; Dahab et al., 2018a; Dahab et al., 2018b; Alnofaie et al., 2016; Aljaloud et al., 2016) use a fixed value of B despite the length of documents. To make B varies with any length of a document, L, and to fix the number of words in a bin, W, the following equation is used:

$$B = 2^{\left\lceil \log\left(\frac{L}{W},2\right) \right\rceil} \quad (1)$$

The equation (1) shows that the length of a document, L, and the desired number of words in a bin, W, are parameters in the function in computing B.

Defining B should consider the average length of the documents and the length of the target context. However, the proposed work evaluates the relatedness of the surrounding terms with a penalty. The penalties depend on distance, such that if the distance increases, the penalty does, too, as will be shown later. A term signal, which is constructed for each term t in a document d, is represented by the following equation, inspired from (Dahab et al., 2016):

$$\tilde{f}_{t,d} = [f_{t,1,d}, f_{t,2,d}, \dots, f_{t,B,d}] \quad (2)$$

where $\tilde{f}_{t,d}$ is the term signal of t in the document d. The expression $f_{t,n,d}$ represents a signal component for $1 \leq n \leq B$. The signal component $f_{t,n,d}$ is equal to 1 if the t is found in the bin n in the document d, otherwise, it is equal to zero.

For the purpose of parallel computing, a file is divided into a number of parts depending on the number of processes, nP, used in the computation. In addition, each process deals with a number of bins, nB.

$$nB = \frac{B}{nP} \quad (3)$$

Each process is assigned an incremental number (i.e. rank) starting from zero till $nP - 1$. The process number k deals with bins from $nB \times k$ till $nB \times (k+1) - 1$. For gathering data from different processes on very efficient space and time, the term signal is represented temporarily as follows:

$$\tilde{f}_{t,d} = [B_i, \dots] \quad (4)$$

where B_i is the bin number in which the term t is found.

3.4. Constructing Concordance Signal

A concordance signal is a sequence of values that shows the occurrence of two given terms together, t_1 and t_2, in all documents in a particular section or bin of a document. The concordance signal for two given terms t_1 and t_2 in a document d is constructed from $\tilde{f}_{t_1,d}$ and $\tilde{f}_{t_2,d}$ using equation (5).

$$C_{t_1,t_2,i,d} = \sum_{k=i-m}^{k=i+m} f_{t_1,i,d} \times \frac{f_{t_2,k,d}}{2^{|k-i|}} \tag{5}$$

where $C_{t_1,t_2,i,d}$ is a single component of the concordance signal. The equation(5) is applied simply after applying the zero-padding technique in both $\tilde{f}_{t_1,d}$ and $\tilde{f}_{t_2,d}$ with m length. The expression $2^{|k-i|}$ represents a penalty if both t_1 and t_2 are close to each other but in different bins. The larger the distance, the larger the penalty, and vice versa. By applying equation (5), the component $C_{t_1,t_2,i,d}$ becomes zero if both terms t_1 and t_2 do not exist in the bin i or if a single term appears while the second term does not. The parameter m is important to enlarge the context, but at the same time, it may affect the accuracy if it is large. The $\tilde{C}_{t_1,t_2,d}$ is a whole signal that is constructed for a single document d. The $\tilde{C}_{t_1,t_2}$ is the summation of all signals $\tilde{C}_{t_1,t_2,d}$ for all documents. The following equation (6) shows how to construct a single component i of $\tilde{C}_{t_1,t_2}$.

$$C_{t_1,t_2,i} = \sum_{d \in D} C_{t_1,t_2,i,d} \tag{6}$$

where D is the set of all documents in the considered corpus.

The concordance signal is constructed for any given input term with all terms in the MCL list to construct a word vector.

3.5. Applying Wavelet-Transform

Wavelets are defined by the wavelet function $\psi(t)$ and scaling function $\varphi(t)$ in the time domain. The wavelet function is described as $\psi \in L^2R$ with a zero average and norm of 1. A wavelet can be scaled and translated by adjusting the parameters s and u, respectively (Daubechies, 1996).

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}) \tag{7}$$

Pywavelet (Lee et al., 2006) is adopted to apply one of the available discrete wavelet-transform algorithms such as Biorthogonal, Coiflets, Daubechies, Symlets, Haar, etc.

The wavelet-transform algorithm is applied on the concordance signal $\tilde{C}_{t_1,t_2}$. The score, which represents the relatedness between two terms, is computed from the magnitude of the signal $\tilde{C}_{t_1,t_2}$.

3.6. Parallel Computing

Most of the work is implemented on Aziz super [1]computer using the Python programming language (3.6.8 Anaconda, Inc.) and MPI-2 (Dalcin et al., 2011). All jobs were submitted using 16 nodes, 7 cores each.

The parallel computing is essential for the following tasks:

- Determining the terms' positions within bins in all fields.
- Searching for a term in all bins.
- Generating concordance signals that are constructed with a given input term and selected terms from SUMO.

[1] https://www.hpcc-kau.com/

4. The Experiments and Results

In this research, many corpora are used, these corpora are listed below:

- The UMBC webBase corpus (http://ebiq.org/r/351http://ebiq.org/r/351). The corpus contains 408 files which is about 48GB.
- The remaining corpora are used through Natural Language Tool Kit (NLTK)[2]:
  - The Gutenberg corpus, which contains some 25,000 electronic books, hosted at http://www.gutenberg.org.
  - The Webtext corpus, which is a collection of web text, includes content from a Firefox discussion forum, conversations overheard, movie script, personal advertisements, and other reviews.
  - The Brown corpus, developed by Brown University, contains text from 500 sources, and the sources have been categorized by genre, such as news, editorial, etc.
  - Inaugural address corpus, which is actually a collection of 55 texts, one for each presidential address.

4.1. Constructing Concordance Signal

Each document is divided into a number of bins, *B*, depending on the number of words in a bin, *W*. In this experiment, the number of words has been selected to (*W = 7*). The concordance signal is constructed with a given input term with a selected subset from SUMO. In this experiment, the MCL has been constructed from 150 terms that are selected based form six base classes: situation (15 terms), action (47 terms), event (15 terms), time (16 terms), personal (21 terms) and physical entity(36 terms). All selected terms from SUMO are in the stem format using the same stemmer so that the comparison will match the stored terms. Using SUMO enables to expand matching among terms from lexical matching to semantic matching.

Ten terms are selected to investigate the relatedness among them. Table (1) shows the relatedness among the selected terms.

	dog	cat	girl	boy	man	woman	mother	father	mouse	horse
dog	1.0	0.50	0.58	0.65	0.61	0.61	0.63	0.64	∞	∞
cat	0.50	1.0	0.68	0.63	0.68	0.72	0.65	0.76	∞	∞
girl	0.58	0.68	1.0	0.76	0.78	0.73	0.69	0.71	∞	∞
boy	0.65	0.63	0.76	1.0	0.72	0.77	0.77	0.7	∞	∞
man	0.61	0.68	0.78	0.72	1.0	0.83	0.74	0.71	∞	∞
woman	0.61	0.72	0.73	0.77	0.83	1.0	0.78	0.69	∞	∞
mother	0.63	0.65	0.69	0.77	0.74	0.78	1.0	0.74	∞	∞
father	0.64	0.76	0.71	0.7	0.71	0.69	0.74	1.0	∞	∞
mouse	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
horse	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞

Table 1. The relatedness among the selected terms

As shown in the previous table, the relatedness relation is symmetric, and any term is totally related to itself. Two terms, mouse and horse, are not related to any terms according to the current MCL. The distance from the term dog to terms man and woman are equal. Also, the distance from the term dog to terms mother and father are approximately equal. Table (2) shows the frequency of the most frequent terms after stemming.

[2] https://www.nltk.org/

Term	Frequency	Term	Frequency	Term	Frequency
accept	10668	access	10174	connect	10339
account	10653	act	10584	consist	10409
present	16996	show	12546	time	19138
activ	11790	address	10582	continu	17037
approach	10089	area	10637	contribut	9675
author	12309	back	10225	control	13164
base	9505	call	12780	cover	10157
care	12377	chang	10720	creat	9763
offer	12204	place	15636	person	14289
close	12195	collect	12406	depend	12171
complet	11416	concern	11021	day	12916
combin	9608	compar	9675	design	15944
develop	16243	differ	11759	direct	11951
discuss	10721	effect	14522	end	12307
establish	9985	exist	11296	expect	10835
hand	12494	happen	10179	head	10094
hope	10953	improv	11217	includ	11842
inform	10799	interest	13382	level	10513
state	14269	organ	13385	work	18957
life	10029	law	10409	market	11693
face	10475	find	10010	follow	11962

Table 2. The Frequency of The Most Frequent Terms

4.2. Parallel Computing

The parallel computing is essential for both determining and searching a term within bins in all files. Using $W = 7$ with the aforementioned corpora makes the number of bins very large which is stored in about ~2.73GB of a compressed indexed file. The following structure is used to store terms in bins.

File Name	Bin Number	Term

5. Conclusion and Future Work

The classical natural language processing systems deal with words as discrete atomic symbols while using vector representations can overcome some of these obstacles and provide much more information about words.

SUMO may be combined with Mid-Level Ontology (MILO), and this merged version has 27,684 terms divided into 7,772 concepts and 19,912 instances, along with ~ 80,000 axioms.

More complicated MCL may be investigated with different corpora. Investigating the proposed work with a different domain, such as sentiment analysis is highly recommended. Applying the suggested work on different natural languages such as Arabic will enrich the language with good resources. The dimensionality is the major problem in representing the word vector.

Moreover, one of the proposed works is to represent MCL using base classes in SUMO instead of terms. In this case, a component in MCL can be a function of all terms in the base class.

Acknowledgment

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH) { King Abdulaziz City for Science and Technology - the Kingdom of Saudi Arabia { award number (12-inf2751-03). The authors also, acknowledge with thanks Science and Technology Unit, King Abdulaziz University for technical support" as well as the members of Aziz support team and the High-Performance Computing Center (HPCC) for the continuous support and guidance.

References

Aljaloud, H., & Dahab, & M. Kamal, M. (2016). Stemmer impact on Quranic mobile information retrieval performance Stemmer impact on quranic mobile information retrieval performance. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 7(12), 135-139.

Al-Mofareji, H., & Kamel, M. ,& Dahab, MY. (2017). WeDoCWT: A new method for web document clustering using discrete wavelet transforms Wedocwt: A new method for web document clustering using discrete wavelet transforms. Journal of Information & Knowledge Management, 16(1), 1-19.

Alnofaie, S., & Dahab, M., & Kamal, M. (2016). A novel information retrieval approach using query expansion and spectral-based A novel information retrieval approach using query expansion and spectral-based. information retrieval,7(9),364-373.

Chen, C., & Gao, S., & Xing, Z. (2016). Mining analogical libraries in q&a discussions-incorporating relational and categorical knowledge into word embedding. 2016 IEEE 23rd international conference on software analysis, evolution, and reengineering (SANER), (1), 338-348.

Costa, A., & Melucci, M. (2010). An information retrieval model based on discrete fourier transform. Information Retrieval Facility Conference Information retrieval facility conference, 84-99.

Cummins, R., & O'Riordan, C. (2009). Learning in a pairwise term-term proximity framework for information retrieval. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 251-258.

Dahab, MY., & Alnofaie, S., & Kamel, M. (2018a). A tutorial on information retrieval using query expansion A tutorial on information retrieval using query expansion. Intelligent Natural Language Processing: Trends and Applications,740, 761-776. Springer.

Dahab, MY., & Kamel, M., & Alnofaie, S. (2016). Further investigations for documents information retrieval based on DWT. International Conference on Advanced Intelligent Systems and Informatics,533, 3-11. Springer.

Dahab, MY., & Kamel, M., & Alnofaie, S. (2018b). An Empirical Study of Documents Information Retrieval Using DWT. Intelligent Natural Language Processing: Trends and Applications Intelligent natural language processing: Trends and applications, 740, 251-264. Springer.

Dalcin, LD., & Paz, RR., & Kler, PA., & Cosimo, A. (2011). Parallel distributed computing using Python Parallel distributed computing using python. Advances in Water Resources, 349,1124-1139.

Daubechies, I. (1996). Where do wavelets come from? A personal point of view. Proceedings of the IEEE, 844, 510-513.

Diwali, A., & Kamel, M., & Dahab, M. (2015). Arabic text-based chat topic classification using discrete wavelet transform. International Journal of Computer Science Issues (IJCSI),12(2), 86-94.

Drozd, A., & Gladkova, A., & Matsuoka, S. (2016). Word Embeddings, Analogies, and Machine Learning: Beyond king – man + woman = queen. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical, 3519-3530. Osaka, JapanThe COLING 2016 Organizing Committee.

Lee, G., & Wasilewski, F., & Gommers, R., & Wohlfahrt, K., & O'Leary, A., & Nahrstaedt, H. (2006). PyWavelets: Wavelet Transforms in Python. Pywavelets: Wavelet transforms in python.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings Dependency-based word embeddings. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers) , 302-308.

Mikolov, T., & Chen, K., & Corrado, & G. Dean, J. (2013a). Efficient estimation of word representations in vector space Ecient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., & Le, QV., & Sutskever, I. (2013b). Exploiting similarities among languages for machine translation Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology Towards a standard upper ontology. Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001, (2-9).

Park, LA., & Palaniswami, M. & Ramamohanarao, K. (2005a). A novel document ranking method using the discrete cosine transform A novel document ranking method using the discrete cosine transform. IEEE transactions on pattern analysis and machine intelligence, 27(1), 130-135.

Park, LA., & Ramamohanarao, K., & Palaniswami, M. (2005b). A novel document retrieval method using the discrete wavelet transform A novel document retrieval method using the discrete wavelet transform. ACM Transactions on Information Systems (TOIS), 23(3), 267-298.

Porter, M. (2008). The Porter stemming algorithm, (2005). Retrieved 12 October 2020 from URL http://www.tartarus.org/martin/PorterStemmer/index. html.

Reiter, N. (2007). Towards a Linking of FrameNet and SUMO. Doctoral dissertation, Master's thesis, Saarland University.

Rong, X. (2014). word2vec parameter learning explained word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

Senel, LK., & Utlu, I., & Yucesoy, V., & Koc, A. & Cukur, T. (2018). Semantic structure and interpretability of word embeddings. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), 1769-1779.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J. Mai, K. 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and GoogleWord2Vec model. International Journal of Geographical Information Science, 31(4), 825-848.