Quantum-like Generalization of Complex Word Embedding: a lightweight approach for textual classification

Amit Kumar Jaiswal[0000−0001−8848−7041], Guilherme Holdack[0000−0001−6169−0488], Ingo Frommholz[0000−0002−5622−5132], and Haiming Liu[0000−0002−0390−3657]

University of Bedfordshire, Luton, UK

Abstract. In this paper, we present an extension, and an evaluation, to existing Quantum like approaches of word embedding for IR tasks that (1) improves complex features detection of word use (e.g., syntax and semantics), (2) enhances how this method extends these aforementioned uses across linguistic contexts (i.e., to model lexical ambiguity) - specifically Question Classification -, and (3) reduces computational resources needed for training and operating Quantum based neural networks, when confronted with existing models. This approach could also be latter applicable to significantly enhance the state-of the-art across Natural Language Processing (NLP) word-level tasks such as entity recognition, part-of-speech tagging, or sentence-level ones such as textual relatedness and entailment, to name a few.

Keywords: Word embedding · Quantum Theory · Word-Context.

1 Introduction

Word embedding [4, 5] is a general technique for treating words as a vector of real valued numbers. It is a well-known distributed form of word representation [6], encoding semantic as well as linguistic information and instruction of words, which generated state-of-the-art results in several Information Retrieval (IR) and NLP tasks in recent times. Although existing research [22, 23] presents architectures proven to be successfully used in several posterior tasks, only few studies exist that analyze the word-embedding mechanism itself, and how enhancing it, or even simplifying it, could lead to better results to existing methods. Particularly, this paper presents an analysis on how simple word embedding optimization can lead to expressive better results on Question Classification task. Specifically, how sensibly reducing word embedding representations from larger pre-trained [13, 8, 14] corpus can prove beneficial to Quantum based models, where complexity and inputs size are factors of great concern.

Being extensively used to better capture cognitive behavior in different domains [1, 2], Quantum mechanics principles can also be applied to IR when it

comes to language related tasks. One example can be outlined on determining textual entailment, or combination [3], of terms. For instance, modern approaches can assign high probabilities to the words '*strong*' and '*coffee*'' in a term '*strong coffee*' if they repeatedly co-occur in a training corpus. However it can lapse to capture the fact that they might occur in an opposite sense - '*Coffee is not very strong*'. By applying Quantum cognitive aspects to word embedding [10], it is stated that users do not superimpose a single polarity or sentiment to each word, where a term subscribes to the global polarity of aggregated words based on the other entities it is coupled with. This resembles the action of tiny particles which remain in all possible states at the same time and hinder each other giving rise to new states based on the relative aspects: words that occur in similar contexts tend to have similar meanings.

By highlighting the importance of the two aforementioned trending concepts, namely word embedding and Quantum cognitive models, in this paper we focus on providing a lightweight method for encouraging researchers to continually embrace new opportunities in IR field when solving textual related tasks. We structure our work in this paper as follows: first we gently introduce a background section on how words in a corpus are trained in form of embeddings (vectors) to compose a vector space. We then bring to light a background on Quantum inspired models for IR and their applications for textual tasks, followed by how trending pre-trained models can represent a potential problem on computing resources when dealing with complex architectures like in deep neural networks [9], and our proposition on how to solve it. On the subsequent section, we present accuracy evaluations on Question Classification task confronting existing literature as a baseline, and then conclude outlining the open research opportunities for this paper.

2 Background

With the advent of Word2Vec, an iterative group of algorithms that record co-occurrence of words at a time rather than capturing all co-occurrence counts explicitly like in singular value decomposition, word embedding technique places words into space in which it approximates (1) spatial distance for computation (2) constant relationships as vectors in space. Recent studies on word representations based on those vectors proposed more efficient and accurate morphological language structures [15, 8] as a natural evolution of this open research field. Global vectors of word representation (GloVe) [8] is an unsupervised learning algorithm for retrieving vector representations for words, trained on assembled global word-to-word co-occurrence stats from a corpus, and the resulting representations shows impressive performance in several NLP and IR tasks [26–29] with linear sub-structures of the word vector space.

However, shifting from word-level to recently proposed character-level models, like FastText in [13], made it efficient to tackle languages peculiarities or even the problem of '*unknown words*', i.e. rare words that were not available in the corpus of the training process of the embedding models. FastText [13] is an

enhanced version of traditional Word2Vec [15, 16], which enriches vocabulary analysis with the usage of character n-grams, i.e., a word is the result of the sum of the many character n-grams that composes it.

3 Related Work

3.1 Quantum like assertion

Several embedding models based on the formalism of Quantum mechanics have been investigated to show dependencies between words and text portions [18–20], which have been modeled as Quantum mixed states and then described by a so called density matrix with its off-diagonal elements depicting word relationship in a Quantum aspect. As opposed to traditional embeddings allocated in one-dimensional vector space, those complex representations are placed in one infinite-dimensional space, called the Hilbert space. Recent advancement of Quantum-inspired models for Information Retrieval tasks studies the non-classical behavior of word dependency relations. Most of these models facilitate the space arena to operate in a real-valued Hilbert space $\mathtt{R}^n$, describing word or a text portion being a real-valued vector or matrix, generally because the paucity of appropriate textual patterns contributes to the imaginary part. Earlier studies [24, 25] confirm that Quantum events cannot be simply explicated without complex numbers, as most models are theoretically narrowed. More related to IR, QWeb [21] is a Quantum theoretical framework for modeling document collections in which the notion of a term is replaced by "whole meaning" that can be a concept or concept mixture described as a state in a Hilbert Space and a superposition of the concept states respectively. In this framework, the complex phases of all concepts have a natural association to the scope of interference among concepts. However, this framework has not yet turned up with its applicability to any IR or NLP tasks as per the authors' knowledge. The Quantum Information Access framework discussed in [17] uses term vectors and their combinations (by means of mixtures and superpositions) to represent information needs in a real-valued Hilbert space. This highly expressive representation of documents and queries is shown to perform similar to established ranking functions like BM25.

In search of a potential Quantum-inspired model, [10] interprets words with complex weights in a linear sequence of latent concepts, and multiple words as a complex sequence of word states, being terms represented either in a mixed or superposition state, which consents with [21] except by the fact that their assumption of terms are described as "entities of meaning" in QWeb. Our experiment is based on sentence-level interpretation and considers a sentence as a blend of words, where a word is described as a Quantum state consisting of two parts, (1) amplitudes of co-occurred words to capture the low-level information and (2) complex phases to depict the emergent meaning or polarity when a word aggregates with other words. In this manner, the meaning from word combination will obviously prevail in the interference among words and will consequently be recorded tacitly in the density matrix representation. This paper conducts a

more exhaustive evaluation of the Quantum model described in [10] that contributes to the domain of both word embedding and Quantum-inspired IR, which can be illustrated as an enhanced evaluation of classical and non-classical embedding approaches, observed as a research study for Quantum-inspired language models and this benchmarking on Complex word embedding can be applicable for QWeb [21] onto developing an application context.

3.2 Embeddings and their Dimensionality Reduction

On the Classical Mixture approach of sentences proposed by [10], the need of representing each word vector $\boldsymbol{w}$ into $\boldsymbol{w}^T\boldsymbol{w}$ (equivalently, in Dirac's notation $|w\rangle$ and $|w\rangle\langle w|$, respectivelly) increases the resources needed for training a model, specially deep models like a neural networks. Publicly available pre-trained word embedding models [14, 13] were used as the foundation for the experiments input. However, such pre-trained models contain embeddings with 300-units dimensions per term, resulting in complex embedding matrices of 300 columns and rows.

The high dimensionality of Quantum structured objects tends to result in lower batch sizes and increased memory and processor usage, consequently taking longer for a network to be trained, even on well-equipped hardware, as no longer the inputs $|w\rangle$ of dimension $(1, ||w\rangle|)$, but $(||w\rangle|, ||w\rangle|)$. Motivated by this challenge to cope with high dimensionality, we include further experiments on embeddings dimensionality reduction, reproducing [11, 12], which, in a combined manner, benchmarks different algorithms but none related to Quantum techniques applied to IR. In [12], by subtracting the 'mean energy' of the vectors contained in the model, hence increasing discrimination between vector indexes, we apply Principal Component Analysis (PCA) technique, to identify the top components responsible for the major variance ratio between each of the units of the word vectors of the whole model vocabulary. Then, by eliminating those top components, and performing further compression (PCA transformation do half the size of the dimensions) on the pre-trained models, it is possible to still outperform original ones at the same time as optimizing training process and time for Quantum complex architectures, as shown in section 4.2.

Although simply extending [12] with the method proposed by [11] would be straightforward, we also address an extension to the latter, with our proposed dynamic way of determining the number of components to be removed from the model based on a threshold. One open line of research in [11] was replacing a fixed threshold of 7 components to be nullified by PCA analysis. Our empirical analysis has shown that removing the components responsible for 20% of the variance, which represents in practice either 6 or 7 components depending on which word embedding model is being analyzed, can still present useful results with the model transformation techniques, independently of which algorithm the embedding model has been trained upon. Below, the two steps of the complete operation are described, including the proposed dynamically generated factor γ in Algorithm 1 , which serves as a counter to identify the number of components to be removed, and the main transformation procedure in Algorithm 2, as originally proposed in [11].

Algorithm 1 Dynamic Model Discrimination

procedure DISCRIMINATE-MODEL(V, threshold = 0.2) ▷ the proposed threshold
  μ = MEAN(V)
  for `n = 1, ..., |V|` **do**
    $\hat{V}$[n] = V[n] - μ
  end for
  $p_{1...d}$ = PCA(V)
  $\gamma = 0$
  partial_ratio = 0.0
  for p in p_i **do**
    γ += 1
    if partial_ratio + p.variance_ratio $\geq$ threshold **then**
      break
    end if
    partial_ratio += p.variance_ratio
  end for
  for `n = 1, ..., |V|` **do**
    V[n] = $\hat{V}$[n] - $\Sigma_{i=1}^{\gamma}(p_i^T V[n])p_i$
  end for
  return V
end procedure

where V = the model being transformed, represented by key/value combination of term and its vector representation, threshold = threshold for the major variance ratio interval

Algorithm 2 Model Reduction

procedure REDUCE-MODEL(V)
  n = ($|V|$)/2
  V = DISCRIMINATE-MODEL(V)
  V = PCA-TRANSFORM(V, n)
  V = DISCRIMINATE-MODEL(V)
  return V
end procedure

where V = the model being transformed, represented by key/value combination of term and its vector representation

4 Evaluation

The goal of our evaluation is to look at performance of the different embedding models and datasets on a Question Classification task. The relevance of this task can be defined by the fact that, traditionally in IR, this represents one problem which focuses on identifying characteristics needed to answer a potential question, or even to determine the type of a question itself, like a 'who', 'how' or 'when' question to be answered by a system [30]. In the following we describe the datasets and the results of our experiment.

4.1 Dataset

For our experiments on Question Classification task to be conducted we made use of datasets on two different stages. The first one was the choice of pre-trained word embedding models, which we will also refer to as '*datasets*' in this section. The second and last step, was determining which dataset, properly saying, would be considered to test our findings. Table 1 lists the models we elected as input for the Complex Word Embeddings generation. For this matter, we elected GloVe [8] and Facebook Research Lab's recently proposed FastText [13].

The motivation behind the selection of these two distinct aforementioned methods is that the first behaves in a similar way of traditional word2vec algorithm, treating each word as one exclusive vector representation, whereas FastText is built on top of the idea that a word is represented by the sum of the different character n-grams it is composed of, *relaxing* the constraints of dealing with rare and unknown words on training phase, and behaving like this in a possible multi-language scenario, since it is agnostic to rigid rules on morphological compositions.

Identification	**Vocabulary Size**	**Corpus size in tokens**
(A) - Wikipedia.Gigaworld.GloVe.6B.300d †	400 thousand	6 Billion
(B) - crawl-300d-2M-vec ‡	2 Million	600 Billion
(C) - GloVe.Common Crawl.840B.300d †	2.2 Million	840 Billion

Table 1. The pre-trained word embedding models selected for this experiment, where †= GloVe algorithm embeddings, ‡= Fasttext algorithm embeddings. All the models have vector dimensions set to 300 units.

Models **(B)** and **(C)** have been also transformed using the techniques described in section 3.2, leading us to 5 word embedding models, properly saying - being 2 versions of each ('transformed' with vector length of 150, and 'original' with vector length of 300).

Table 2, on the other hand, shows the structured datasets that were used to carry out the experiments. Namely, TREC-10 Dataset for Question Classification, and Stanford Sentiment Treebank, which were chosen as a way of fairly comparing existing baseline expressed in [10].

4.2 Experiments and Results

To compare the performance of the different embedding models and the datasets explored in Table 3, we here present one concise table contemplating our modified embedding models executed in classic fashion (words represented as real valued numbers - datasets appended with 'R'), vs the reproduced mixture model

Identification	Number of records	Number of classes
TREC 10 Question Classification	5,952	6
SST-2	70,042	2

Table 2. Datasets description

from [10] (datasets appended with 'M'). We observe how the character n-grams based FastText **(B)** model outperforms mostly all other models when it comes to traditional word embedding input, and on the other hand, how GloVe based models produce higher accuracy when applying the Complex Word Embeddings technique. In particular, it is visible how the reduced versions of the pre-trained embedding models outperform original ones with a considerable distance, or almost perform equally with minimal percentage difference. This comes with a great advantage, since in practice it represents a model reduced in 50% of size (less resources and time needed for training), and also lower complexity (150 units vs 300 of pre-trained versions). Nevertheless, it must also be highlighted how the 50% reduction of model **(C)**, which was trained in a huge corpus of 840 Billion Tokens, does not considerably degrade the overall performance if compared to its original version. It becomes a term on the equation for an expert to decide the trade-off between model size and complexity *versus* performance and small degradation in accuracy.

Model	SST-2-R	TREC-R	SST-2-M	TREC-M
(A)	78.47 %	79.80 %	82.14 %	75.48 %
(B)	**79.29 %**	79.30 %	81.73 %	84.84 %
(B) - reduced	78.86 %	**82.30 %**	81.83 %	85.20 %
(C)	79.10 %	80.50 %	**82.42 %**	84.20 %
(C) - reduced	78.42 %	80.00 %	82.21 %	**85.48** %

Table 3. Experiments conducted. Here, the metrics for evaluation is the accuracy on the classification. In bold, the best results for each method.

4.3 Training steps - time reduction

As one important aspect on the dimensionality reduction of the pre-trained word embedding models, we also plot in Table 4 the difference on the scale of 'time

per epoch', listing in **seconds** how much time each model took, per epoch, to be trained on the different tasks:

Model	SST-2-R	TREC-R	SST-2-M	TREC-M
(A)	31s	8s	974s	51s
(B)	31s	8s	974s	51s
(B) - reduced	20s	5s	332s	18s
(C)	31s	8s	972s	51s
(C) - reduced	20s	5s	332s	18s

Table 4. The times per epoch relation presents the difference on training time with equivalent accuracy as original pre-trained models, however with 50% of the original size of each.

We also place as publicly available the gists containing the logs of execution of each of the four identified tasks, namely SST-2-R[1], TREC-R[2], SST-2-M[3], TREC-M[4].

5 Conclusion and Future Work

The proposed reduction accomplishes better and more efficient execution than the state-of the-art Quantum and classical models evaluated on the Question Classification and Sentiment Analysis datasets, with the usage of large pre-trained English corpus based on different word embedding techniques. As future work, we also see an opportunity for an analysis on how to map the impact, or weight, of each word in a sentence, leveraging performance of learning tasks which could then lead to a bigger model overall accuracy. This open research could lead to establishing an interesting real-valued factor γ that could increase or decrease the importance of a given term in a sentence, according to the relevance the world represents to a context.

ACKNOWLEDGMENTS

This work is supported by the Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Unions Horizon 2020

[1] https://bit.ly/2Oq1WR1
[2] https://bit.ly/2Ae72wx
[3] https://bit.ly/2LOw0qG
[4] https://bit.ly/2LrFQzs

research and innovation programme under the Marie Sklodowska-Curie grant agreement No 721321.

References

1. Busemeyer, J. R., & Bruza, P. D. (2012). Quantum models of cognition and decision. Cambridge University Press.
2. Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling?. Behavioral and Brain Sciences, 36(3), 255-274.
3. Bruza, P. D., Kitto, K., Ramm, B. J., & Sitbon, L. (2015). A probabilistic framework for analysing the compositionality of conceptual combinations. Journal of Mathematical Psychology, 67, 26-38.
4. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb), 1137-1155.
5. Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167). ACM.
6. Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 384-394). Association for Computational Linguistics.
7. Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 238-247).
8. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
9. Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J. F., ... & Pal, C. J. (2017). Deep complex networks. arXiv preprint arXiv:1705.09792.
10. Li, Q., Uprety, S., Wang, B., & Song, D. (2018). Quantum-inspired Complex Word Embedding. arXiv preprint arXiv:1805.11351.
11. Raunak, V. (2017). Effective Dimensionality Reduction for Word Embeddings. arXiv preprint arXiv:1708.03629.
12. Mu, J., Bhat, S., & Viswanath, P. (2017). All-but-the-top: simple and effective postprocessing for word representations. arXiv preprint arXiv:1702.01417.
13. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.
14. Google Word2Vec Pre-Trained Models. Available at: https://code.google.com/archive/p/word2vec/, June 2018
15. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
17. Piwowarski, B., Frommholz, I., Lalmas, M. & van Rijsbergen, Keith (2010, October). What can Quantum Theory Bring to Information Retrieval? In Proc. 19th International Conference on Information and Knowledge Management (CIKM 2010) (pp. 59–68). ACM.

18. Sordoni, A., Nie, J. Y., & Bengio, Y. (2013, July). Modeling term dependencies with quantum language models for IR. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 653-662). ACM.
19. Xie, M., Hou, Y., Zhang, P., Li, J., Li, W., & Song, D. (2015). Modeling quantum entanglements in quantum language models.
20. Zhang, P., Niu, J., Su, Z., Wang, B., Ma, L., & Song, D. (2018). End-to-End Quantum-like Language Models with Application to Question Answering.
21. Aerts, D., Argulles, J. A., Beltran, L., Beltran, L., Distrito, I., de Bianchi, M. S., ... & Veloz, T. (2017). Towards a Quantum World Wide Web. arXiv preprint arXiv:1703.06642.
22. Ganguly, D., Roy, D., Mitra, M., & Jones, G. J. (2015, August). Word embedding based generalized language model for information retrieval. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (pp. 795-798). ACM.
23. Kumar, A., & Soman, K. P. (2016). Amritacen at semeval-2016 task 11: Complex word identification using word embedding. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 1022-1027).
24. Kadin, A. M. (2005). Quantum Mechanics without Complex Numbers: A Simple Model for the Electron Wavefunction Including Spin. arXiv preprint quant-ph/0502139.
25. Kwong, C. P. (2009). The mystery of square root of minus one in quantum mechanics, and its demystification. arXiv preprint arXiv:0912.3996.
26. Salle, A., Idiart, M., & Villavicencio, A. (2016). Enhancing the LexVec Distributed Word Representation Model Using Positional Contexts and External Memory. arXiv preprint arXiv:1606.01283.
27. Salle, A., Idiart, M., & Villavicencio, A. (2016). Matrix factorization using window sampling and negative sampling for improved word representations. arXiv preprint arXiv:1606.00819.
28. Galke, L., Saleh, A., & Scherp, A. (2017). Word Embeddings for Practical Information Retrieval. INFORMATIK 2017.
29. Diaz, F., Mitra, B., & Craswell, N. (2016). Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891.
30. Siddhartha1, M., Singh2, A.K., Dwivedi3 S.K. (2017). Question Analysis and Classification for Question Answering System. International Journal of Innovative Research in Computer and Communication Engineering (Vol. 5, Issue 9).