# Impact of Document Representation on Neural Ad hoc Retrieval

Ebrahim Bagheri
Ryerson University
bagheri@ryerson.ca

Faezeh Ensan
Ferdowsi University of Mashhad
ensan@um.ac.ir

Feras Al-Obeidat
Zayed University
Feras.al-obeidat@zu.ac.ae

## ABSTRACT

Neural embeddings have been effectively integrated into information retrieval tasks including ad hoc retrieval. One of the benefits of neural embeddings is they allow for the calculation of the similarity between queries and documents through vector similarity calculation methods. While such methods have been effective for document matching, they have an inherent bias towards documents that are sized relatively similarly. Therefore, the difference between the query and document lengths, referred to as the *query-document size imbalance* problem, becomes an issue when incorporating neural embeddings and their associated similarity calculation models into the ad hoc document retrieval process. In this paper, we propose that *document representation* methods need to be used to address the size imbalance problem and empirically show their impact on the performance of neural embedding-based ad hoc retrieval. In addition, we explore several types of document representation methods and investigate their impact on the retrieval process. We conduct our experiments on three widely used standard corpora, namely Clueweb09B, Clueweb12B and Robust04 and their associated topics. Summarily, we find that document representation methods are able to effectively address the query-document size imbalance problem and significantly improve the performance of neural ad hoc retrieval. In addition, we find that a document representation method based on a simple term-frequency shows significantly better performance compared to more sophisticated representation methods such as neural composition and aspect-based methods.

**ACM Reference Format:**
Ebrahim Bagheri, Faezeh Ensan, and Feras Al-Obeidat. 2018. Impact of Document Representation on Neural Ad hoc Retrieval. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269314

## 1 INTRODUCTION

The low-dimensional and dense vector representation of terms based most recently on neural network models, known as *neural (word) embeddings*, has received growing attention within the information retrieval community. This representation model has shown to have desirable syntactic, semantic and geometric properties and is computationally practical to be used in real-world applications [15]. As such, neural embeddings have been incorporated into tasks such as language modeling [7], and query expansion [20], just to name a few. Ad hoc document retrieval literature has also explored the possibility of incorporating neural embeddings into the retrieval process. One of the focus areas of research in this domain has been to study the impact of the quality and type of embeddings on retrieval performance. The work by Zuccon et al. [21] was among the earlier work to systematically incorporate neural embeddings within the context of a translation language model and show that retrieval could be enhanced even if the embeddings were trained based on a completely separate corpus. Zamani and Croft [19] showed that neural embeddings trained on a *global* non-aligned corpus can improve query expansion and retrieval effectiveness.

In more recent work, researchers have considered the distinction between globally trained and locally trained neural embeddings and its impact on the retrieval process. For instance, Kuzi et al. [12] focus on training neural embeddings based on the document collection on which the queries will be executed. They found that *locally* trained embeddings, when interpolated with an effective retrieval model such as RM3 lead to a different set of more effective expansions terms that are complementary to the baseline and lead to increased retrieval effectiveness.

Neural embeddings, either trained locally or globally, can be incorporated into query likelihood (QL) models to provide a measure of similarity or relevance between query and document pairs, referred to as *relevance functions* and denoted by $\phi$. More succinctly, assuming a uniform prior and that $P(q_i|D)$ can be estimated by the relevance function $\phi$, query likelihood can be defined as:

$$P(D|Q) \overset{\text{rank}}{\simeq} \prod_{q_i \in Q} \phi(q_i, D). \quad (1)$$

where $q_i$ is a query term and $D$ is the document. In this context, there have been strong work that provide various types of implementation for $\phi(q_i, D)$ based on term and/or semantic entity similarity [5, 9]. Within the context of neural embeddings, $\phi(q_i, D)$ can be viewed as the similarity of the vector representations for $q_i$ and $D$ that are trained on local or global corpora. A more systematic approach could be to generalize $\prod_{q_i} \phi(q_i, D)$ into $\Phi(Q, D)$, which is a relevance function for computing the similarity between $Q$ and $D$ depending on the vector representation of their constituting terms in $Q$ and $D$, i.e., $q_i$ and $d_i$, respectively. Kusner et al. [11] have suggested that a suitable method for calculating the similarity of two embedding vector collections, e.g., a query (Q) and a document (D), is the minimum cumulative distance of the best matching embedding pairs in the two documents.

There have been recent empirical studies based on Kusner's work that show ad hoc retrieval based solely on neural embeddings cannot perform competitively with traditional keyword-based techniques such as the sequential dependence model (SDM) [1]. For instance, it was shown that when using pre-trained global neural embeddings, the performance of ad hoc retrieval can be as


*CIKM '18, October 22–26, 2018, Torino, Italy*

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269314

much as 40% and 60% worse compared to SDM on ClueWeb09B and ClueWeb12B corpora, respectively. One of the reasons for the poor performance is the need for the two embedding collections, one representing the query and one the document, to be sized relatively similarly. However, when dealing with calculating the similarity between a query and a document, i.e., $\Phi(Q, D)$, in most cases, if not all, the size of the document is significantly larger than the size of the query. Some studies have reported that the average search query length is 2.4 terms while documents often consist of hundreds of terms. As such and based on [11], the terms in the query would connect to the best matching terms in the document, for which a variety of options are available given the large size of terms in the document. On the other hand, the terms in the document would only have a limited number of options for being matched, i.e., equivalent to the number of terms in the query.

Given an increasing number of recent neural retrieval models such as [8, 10, 16] are based on the distance of embedding sets, it is essential to systematically explore how the query-document size imbalance problem can be addressed. The size imbalance problem is not as noticeable in keyword-based retrieval due to the requirement for matching common, ordered or un-ordered, n-grams across query and document spaces. However, in neural retrieval and due to a *soft matching* strategy, the size imbalance issue is an important consideration. Our work in this paper is among the first to address the size imbalance problem by proposing to apply *document representation* techniques in the context of neural ad hoc retrieval. On this basis, the objective of our work in this paper is two-folds: (1) To show how document representation techniques can be used to address the query-document size imbalance problem and the extent to which such a strategy can improve the performance of neural ad hoc retrieval; (2) To systematically compare the impact of different document representation techniques on the performance of neural ad hoc retrieval and determine whether there is a significant difference between such techniques.

In order to conduct systematic evaluation, we performed experiments on three widely used document collections, namely Clueweb09B, ClueWeb12 and Robust04 based on TREC topics and their relevance judgments. We also adopted document representation techniques based on three non-overlapping principles related to (1) term frequency, (2) neural composition, and (3) document aspects. Summarily, we found that regardless of the document representation method used, the performance of neural ad hoc retrieval shows statistically significant improvement when a document representation technique is employed. Furthermore, our experiments showed that a simple term frequency based document representation, while quite inexpensive to compute, leads to the best performance on neural ad hoc retrieval and is statistically significant over other document representation techniques.

## 2 EXPERIMENTAL METHODOLOGY

**Corpora:** In our experiments, we used ClueWeb09 Category B dataset (ClueWeb09B), which consists of the first 50 million English Web pages of ClueWeb09, ClueWeb12 Category B (ClueWeb12B) dataset, which is a subset of over 50 million documents from the ClueWeb12 corpus as well as Robust04, which is a collection of over 500 thousand documents from Financial Times, the Federal Register 94, the LA Times, and FBIS. As proposed in [5, 18], we pooled the top-100 documents retrieved by SDM [14] per query, which were then re-ranked based on the work in this paper. Similar to [18], we used the SDM runs provided by [3], which is a well-tuned Galago-based implementation.

**Topics:** We used the standard TREC topics related to each of these corpora for the experiments. For ClueWeb09B: TREC topics 1-200, ClueWeb12B: TREC topics 201-250 and Robust04: TREC topics 301-450 and 601-650 were used.

**Neural Embeddings:** The neural embeddings used in this paper were the pre-trained word vectors learnt using GloVe [17] over the Common Crawl dataset with a dimension size of 300 and vocabulary size of 2.2 million terms (http://nlp.stanford.edu/data/glove.840B.300d.zip). The advantage of this embedding collection is that it is trained based on actual crawled Web pages and hence closely resembles an ad hoc retrieval scenario.

**Evaluation Metrics:** Retrieval effectiveness was evaluated with standard metrics including Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain at 20 (NDCG@20).

**Base Retrieval Model:** The method proposed in [11] serves as the core foundation for several recent neural retrieval methods [8, 10, 16] and as such will serve as our base neural retrieval model in this paper. This method provides pairwise query-document similarity scores based on the similarity of two collections of neural embeddings, which are used to rank documents for each query. Let us assume that $\mathbb{R}^D$ is some representation of document $D$ produced by some document representation technique. Given $\mathbb{R}^D$, it is possible to compute the similarity between $Q$ and $D$ by computing the distance between $Q$ and $\mathbb{R}^D$ by connecting terms in $Q$ to the best matching terms in $\mathbb{R}^D$. As defined in [11] and within the context of our work, matrix $\boldsymbol{T}$ is a *flow matrix* in which $\boldsymbol{T}_{i,j}$ shows to what degree term $i$ in $Q$ is connected to a term $\mathbb{R}^D_j$ in $D$. Matrix $\boldsymbol{T}_{i,j}$, which essentially determines what '*query term-document representation term*' pairs from the query and document spaces should be connected to each other, needs to be learnt based on a linear optimization program. To this end, the distance between a query and a document can be calculated by minimizing the following linear optimization function based on specialized solvers:

$$\Upsilon(Q, D) = min \sum_{i=1}^{|Q|} \sum_{j=1}^{|\mathbb{R}^D|} \boldsymbol{T}_{i,j} \times d(i, j) \tag{2}$$

where $d(i, j)$ is the distance between query term $i$ and term $j$ from the document representation, denoted as $\mathbb{R}^D_j$. The distance function $d(i, j)$ can be defined as the Euclidean distance between the neural embeddings of $q_i$ and $\mathbb{R}^D_j$. To consider term frequency, the objective function is minimized with the following constraints:

$$\sum_{i=1}^{|Q|} \boldsymbol{T}_{i,j} = |\mathbb{R}^D_j| \tag{3}$$

$$\sum_{j=1}^{|\mathbb{R}^D|} \boldsymbol{T}_{i,j} = f(q_i) \tag{4}$$

where $f(q_i)$ is the normalized frequency of query term $q_i$ and $|\mathbb{R}^D_j|$ is its frequency in the document. Hence, $P(D|Q)$ is estimated based on $\Upsilon(Q, D)$ and used to rank documents for each query.

**Table 1: Performance of the base retrieval model before and after document representation is applied. $^{\blacktriangle}$, $^{\dagger}$, $^{\ddagger}$ indicate statistical significant via paired t-test at 0.05 over Neural Composition, Aspects and Without Document Representation, respectively.**

| | | Term Frequency | Neural Composition | Aspects | Without Document Representation [11] |
|---|---|---|---|---|---|
| MAP | ClueWeb09B | $0.092^{\blacktriangle\dagger\ddagger}$ | $0.074^{\ddagger}$ | $0.071^{\ddagger}$ | 0.05 |
| | ClueWeb12B | $0.038^{\blacktriangle\dagger\ddagger}$ | $0.027^{\ddagger}$ | $0.025^{\ddagger}$ | 0.016 |
| | Robust04 | $0.187^{\blacktriangle\dagger\ddagger}$ | $0.147^{\ddagger}$ | $0.151^{\ddagger}$ | 0.07 |
| NDCG@20 | ClueWeb09B | $0.188^{\blacktriangle\dagger\ddagger}$ | $0.125^{\ddagger}$ | $0.121^{\ddagger}$ | 0.061 |
| | ClueWeb12B | $0.114^{\blacktriangle\dagger\ddagger}$ | $0.074^{\dagger\ddagger}$ | $0.067^{\ddagger}$ | 0.029 |
| | Robust04 | $0.34^{\blacktriangle\dagger\ddagger}$ | $0.234^{\ddagger}$ | $0.26^{\blacktriangle\ddagger}$ | 0.1 |

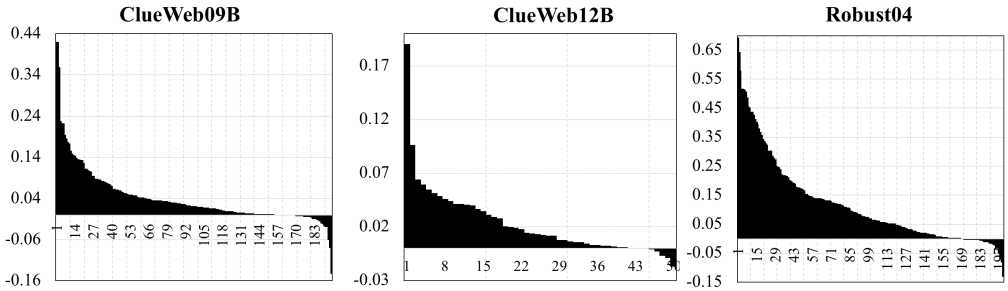


**Figure 1: Delta of MAP for Term Frequency document representation vs base retrieval model.**

**Document Representation:** The objective of document representation techniques is to develop a transformation from $D$ to $\mathbb{R}^D$, which would be a more appropriate representative of $D$ in the context of a certain task. We explore three main document representation techniques in this paper, namely *term-frequency*, *neural composition*, and *aspect-based* methods.

*Term-frequency representation:* The first document representation technique, which is quite inexpensive to compute, selects the top-$k$ terms form $D$ with the highest tf-idf values. As such $\mathbb{R}^D$ based on the tf-idf document representation technique would be a bag of terms with high discriminative power for the document in the corpus because the selected terms are frequent within the document and less frequent across the whole corpus. In our experiments, we performed 10-fold cross validation over the three corpora to determine the best value for $k$, which was deemed to be 10 for all three corpora. This value for $k$ was used in our experiments.

*Neural composition representation:* The second document representation method that we adopt is based on the neural composition of terms within a document, known as *paragraph vectors* that is used to uniquely represent a document or a paragraph within a document [13] through a vector representation. Paragraph vectors are generalizations of word embeddings, i.e., Skipgram and CBOW [15], where an additional vector is learned for any sequence of terms, which represents the missing information of the current context and acts as a memory of the topic of the paragraph. The advantage of the paragraph vector representation is that it takes term order into account the same way as n-gram models would. This is important for our case as prior keyword-based models, such as the potential functions in SDM [14], have shown that n-gram overlaps are strong indicators of query-document similarity. We learn the representation of each document in the corpus based on Gensim's implementation of paragraph vectors[1]. Our experiments did not show statistically significant difference between the PV-DM and PV-DBOW variations of the paragraph vector model and hence we only resort to reporting the PV-DM model in this paper.

*Aspect-based representation:* There have been work in the literature that propose to view a document as a set of underlying *aspects*, which are disjoint components within a document that delineate the main topics within the document [4]. We develop such aspect-based representation of each document by viewing each document in graphical form denoted by $G = (V, E)$ where the nodes $V$ are the embeddings of the terms in the document and the weight of the edges $E$ are the vector similarity of the embedding representation of the terms. Within the document graph, aspects would be identified as highly *modular sub-classes* of nodes [6]. It is possible to find aspects in a document graph without the need for prior knowledge of the number of aspects based on a greedy algorithm with the time complexity of $O(n \log n)$ [2]. A sketch of the greedy algorithm includes the iterative merging of terms from the document graph as long as the degree of modularity increases. Merged nodes into clusters are then considered to be folded into singular nodes based on which the process of merging with other nodes in the document graph is repeated until no more increase in modularity is observed. The aspect-based representation of a document $D$ would consist of $c$ non-overlapping term clusters, each denoted as $\mathbb{R}^D_i = \{r^D_{1,i}, \ldots, r^D_{|\mathbb{R}^D_i|,i}\}$, such that $V = \bigcup_{i=1}^{c} \mathbb{R}^D_i$ and $\mathbb{R}^D_i \cap \mathbb{R}^D_j = \emptyset,\ \forall i, j$. The vector representation of each aspect would be the centroid of its member embeddings. It should be noted that $c$ is determined by the greedy heuristic.

## 3 FINDINGS

We conducted experiments on the three corpora using the base retrieval model with and without the different document representation techniques to explore whether the document representation techniques can address the query-document size imbalance problem and hence impact the performance of neural ad hoc retrieval.

[1] https://radimrehurek.com/gensim/models/doc2vec.html

**Table 2: Help/hurt queries vs no document representation.**

| Corpus | Document Representation | Helped | Hurt | Ratio |
|---|---|---|---|---|
| ClueWeb09B | Term Frequency | 151 | 30 | 5.03 |
| | Neural Composition | 131 | 52 | 2.52 |
| | Aspects | 129 | 51 | 2.53 |
| ClueWeb12B | Term Frequency | 42 | 7 | 6.00 |
| | Neural Composition | 39 | 7 | 5.57 |
| | Aspects | 38 | 8 | 4.75 |
| Robust04 | Term Frequency | 168 | 30 | 5.60 |
| | Neural Composition | 156 | 43 | 3.63 |
| | Aspects | 160 | 38 | 4.21 |

The values of the performance measures are reported in Table 1. The table shows both MAP and NDCG@20 metrics when the document representation techniques have been applied and compares it when no document representation is used. As seen in the table, the application of the document representation techniques has resulted in statistically significant improvement over both MAP and NDCG@20. This shows that the reduction of the document size based on the document representation methods, which addresses the *query-document size imbalance* problem can significantly impact the performance of neural ad hoc retrieval.

Now, we also compare the performance of the different document representation techniques to each other. Our experiments show that while the neural composition and aspect-based representation are computationally expensive, they result in significantly weaker performance compared to the term frequency approach. This has been consistently observed across the three corpora and for both of the evaluation metrics. The improvement gained over both metrics as a result of the term frequency method is quite notable as it performs close to two times better on MAP and three times better on NDCG@20 on the different corpora. It is worth noting that the performance of neural composition and aspect-based methods are both statistical significant over the base retrieval model and similar compared to each other without statistically significant difference.

We further report the number of queries that have been helped or hurt based on the application of the document representation techniques compared to the base retrieval method according to change of MAP in Table 2. As expected, the term frequency method helps the highest number of queries and negatively impacts the lowest number of queries across the three corpora. The ratio of help to hurt queries is very significant in the term frequency based document representation method and consistently shows at least five times more helped queries compared to hurt queries. Figure 1 places the number of helped and hurt queries in context. As seen in the figure, the limited number of hurt queries have only been hurt to a much smaller extent compared to the degree that the other queries have been helped. For instance, in ClueWeb12B, the worst query has been hurt by −0.017, while the maximum helped query is improved by +0.1902, which is ten folds larger. A similar trend can be seen in the other two corpora as well.

## 4 CONCLUDING REMARKS

The work in this paper focused on systematically exploring whether *document representation* techniques can help improve the performance of neural ad hoc retrieval techniques by addressing the *query-document size imbalance* problem. The findings of our work can be summarized as follows:

**(1)** The application of document representation methods can significantly impact the performance of neural ad hoc retrieval methods regardless of the type of the document representation method that is used. Such improvement can be attributed to the need to balance the query and document sizes as assumed by the underlying similarity computation method.

**(2)** From among the document representation methods, the term frequency based representation, which is light weight to compute, shows the most significant improvement over both the base retrieval model as well as the other document representation techniques. The significant improvement can be observed in both the number of helped queries and the degree to which the queries are helped.

Our important finding is that the employment of a light-weight document representation technique, such as term frequency method, helps the performance of neural ad hoc retrieval significantly and as such is an essential pre-processing step in the process of neural ad hoc retrieval.

## REFERENCES


[1] Ebrahim Bagheri, Faezeh Ensan, and Feras Al-Obeidat. 2018. Neural Word and Entity Embeddings for Ad hoc Retrieval. *Information Processing and Management* 54, 2 (2018), 339–357.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *JSTAT* 2008, 10 (2008), P10008.

[3] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. In *SIGIR*. ACM, 365–374.

[4] Wim De Smet and Marie-Francine Moens. 2009. An aspect based document representation for event clustering. In *Proceedings of the 19th Meeting of Computational Linguistics*. 55–68.

[5] Faezeh Ensan and Ebrahim Bagheri. 2017. Document Retrieval Model Through Semantic Linking. In *WSDM 2017*. 181–190.

[6] Dario Fasino and Francesco Tudisco. 2014. An algebraic analysis of the graph modularity. *SIAM J. Matrix Anal. Appl.* 35, 3 (2014), 997–1018.

[7] Debasis Ganguly, Dwaipayan Roy, M. Mitra, and G. Jones. 2015. Word embedding based generalized language model for information retrieval. In *SIGIR*. 795–798.

[8] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. Semantic matching by non-linear word transportation for information retrieval. In *CIKM*. 701–710.

[9] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2016. Exploiting entity linking in queries for entity retrieval. In *ICTIR 2016*. ACM, 209–218.

[10] Sun Kim, Nicolas Fiorini, W. John Wilbur, and Zhiyong Lu. 2017. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. *Journal of Biomedical Informatics* 75 (2017), 122–127.

[11] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML*. 957–966.

[12] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query Expansion Using Word Embeddings. In *CIKM*. 1929–1932.

[13] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*. 1188–1196.

[14] Donald Metzler and W Bruce Croft. 2005. A Markov random field model for term dependencies. In *SIGIR*. 472–479.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. 3111–3119.

[16] Sunil Mohan, Nicolas Fiorini, Sun Kim, and Zhiyong Lu. 2017. Deep Learning for Biomedical Information Retrieval: Learning Textual Relevance from Click Logs. *BioNLP 2017* (2017), 222–231.

[17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*. 1532–1543.

[18] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. 2017. Learning to Attend and to Rank with Word-Entity Duets. In *SIGIR*. 763–772.

[19] Hamed Zamani and W Bruce Croft. 2016. Embedding-based query language models. In *ICTIR 2016*. ACM, 147–156.

[20] Hamed Zamani and W Bruce Croft. 2016. Estimating embedding vectors for queries. In *ICTIR2016*. 123–132.

[21] Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and Evaluating Neural Word Embeddings in Information Retrieval. In *ADCS 2015*.