Hyperspherical Query Likelihood Models with Word Embeddings

Ryo Masumura[†] **Taichi Asami**[†] **Hirokazu Masataki**[†]
Kugatsu Sadamitsu[‡] **Kyosuke Nishida**[†] **Ryuichiro Higashinaka**[†]
NTT Media Intelligence Laboratories, NTT Corporation,
1-1, Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847, Japan
[†] {masumura.ryo, asami.taichi, masataki.hirokazu
nishida.kyosuke, higashinaka.ryuichiro}@lab.ntt.co.jp
[‡] k.sadamitsu.ic@future.co.jp

Abstract

This paper presents an initial study on hyperspherical query likelihood models (QLMs) for information retrieval (IR). Our motivation is to naturally utilize pretrained word embeddings for probabilistic IR. To this end, key idea is to directly leverage the word embeddings as random variables for directional probabilistic models based on von Mises-Fisher distributions that are familiar to cosine distances. The proposed method enables us to theoretically take semantic similarities between document and target queries into consideration without introducing heuristic expansion techniques. In addition, this paper reveals relationships between hyperspherical QLMs and conventional QLMs. Experiments show document retrieval evaluation results in which a hyperspherical QLM is compared to conventional QLMs and document distance metrics using word or document embeddings.

1 Introduction

In information retrieval (IR), language modeling is known to be one of the most successful techniques (Ponte and Croft, 1998). A typical usage is query likelihood models (QLMs), in which language models are constructed from each retrieved document. In QLM-based probabilistic IR, the documents are ranked by probabilities for which a query can be generated by the document language model.

In this field, categorical QLMs which model generative probability of words using categorical distributions are fundamental models (Ponte and Croft, 1998; Zhai and Lafferty, 2001). It is known that the categorical QLMs do not perform well for vocabulary mismatches because categorical distribution cannot consider semantic relationships between words. Therefore, several expansion techniques such as query expansion (Bai et al., 2005), translation QLMs (Berger and Lafferty, 1999), and latent variable models (Wei and Croft, 2006) have been proposed in order to take semantic relationships between document and target query into account.

Recently, word embeddings, which are continuous vector representations embedding word semantic information, have been utilized for enhancing the previous expansion techniques (Zhang et al., 2016; Mitra and Craswell, 2017). The word embeddings can be easily acquired in an unsupervised manner from large scale text sets based on embedding modeling, i.e., skip-gram, continuous bag-of-words (CBOW) (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). Zuccon et al. (2015); Ganguly et al. (2015); Zamani and Croft (2016a) used the word embeddings in order to assist translation QLMs. Zamani and Croft (2016b); Kuzi et al. (2016) used the word embeddings in order to perform query expansion. However, previous word embedding-based probabilistic IR methods have no theoretical validity since the word embeddings were heuristically introduced.

In order to perform more natural word embedding based probabilistic IR, our key idea is to directly leverage word embeddings rather than words as random variables for language models. In fact, the word embeddings can capture semantic similarity of words using directional information based on cosine distance (Mnih and Kavukcuglu, 2013). This motivates us to introduce directional probabilistic models based on von Mises-Fisher distributions which are familiar to the cosine distance (Banerjee et al., 2005; Sra, 2016).

This paper proposes a hyperspherical QLMs in which random variables are modeled by a mix-

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 210–216,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

ture of von Mises-Fisher distributions. The hyperspherical QLMs can theoretically utilize word embeddings for probabilistic IR without introducing heuristic formulations. Main contributions are summarized as follows.

- This paper formulates hyperspherical QLMs based on both a maximum likelihood estimation and a maximum a posteriori estimation.
- This paper reveals that hyperspherical QLMs can be represented as an extended form of categorical QLMs and a theoretical form of translation QLMs.
- This paper shows document retrieval evaluation results in which a hyperspherical QLM is compared with conventional QLMs and document distance metrics with word or document embeddings.

2 Related Work

This paper is closely related to document distance metrics using word or document embeddings. One major distance metric is the cosine distance between two document vectors that are composed by averaging word embeddings (Vulic and Moens, 2015; Brokos et al., 2016) or document embeddings called paragraph vectors (PVs) (Le and Mikolov, 2014). Another highly efficient distance metric is word mover's distance (WMD), which leverages word embeddings (Kusner et al., 2015). In this work, we also examined these distance metrics in a document retrieval evaluation.

Generative models of word embeddings have recently been proposed in topic modeling in order to capture the semantic structure of words and documents (Das et al., 2015; Batmanghelich et al., 2016). To the best of our knowledge, this paper is the first work on language modeling that handles word embeddings as random variables.

3 IR based on QLMs

IR based on probabilistic modeling uses the probability of a document D given a query Q. One of the most famous approaches is QLM-based IR in which documents are ranked by the probabilities that a query can be generated by the document language model (Ponte and Croft, 1998). Given a query $Q = \{w_1, \cdots, w_T\}$, IR based on QLMs ranks documents as

$$P(D|Q) \overset{\text{rank}}{\propto} \prod_{t=1}^{T} P(w_t|\mathbf{\Theta}_D), \tag{1}$$

where $\mathbf{\Theta}_D$ denotes a parameter of QLM for D.

3.1 Categorical QLMs

Categorical QLMs model $P(w|\mathbf{\Theta}_D)$ using a categorical distribution. In a maximum likelihood (ML) estimation for the categorical QLM, a generative probability of a word w is defined as

$$P(w|\mathbf{\Theta}_D^{\texttt{ML}}) = \frac{c(w,D)}{|D|}, \tag{2}$$

where $c(w,D)$ is the word count of w in D, and $|D|$ is the number of all words in D.

In a maximum a posteriori (MAP) estimation, a document collection C in which all of the retrieved documents are included is used for a prior (Zhai and Lafferty, 2001). MAP estimated generative probability of a word w is defined as

$$P(w|\mathbf{\Theta}_D^{\texttt{MAP}}) = \frac{c(w,D) + \tau\frac{c(w,C)}{|C|}}{|D| + \tau}, \tag{3}$$

where $c(w,D)$ is the word count of w in C, and $|C|$ is the number of all words in C. τ is a hyper parameter for adjusting smoothing.

3.2 Translation QLMs

Translation QLMs were introduced for expanding categorical QLMs (Berger and Lafferty, 1999). The translation QLMs are usually used together with the categorical QLMs, and enable us to take into account relationships between a word in the query and semantically related words in the document. A generative probability of a word w is defined as

$$P(w|\mathbf{\Theta}_D^{\texttt{TR}}) = \sum_{v\in\mathcal{V}} P(v|\mathbf{\Theta}_D)P(w|v), \tag{4}$$

where $\mathcal{V}$ is the vocabulary. $P(v|\mathbf{\Theta}_D)$ is the generative probability of a word v, which is also calculated by Eq. (2) or (3). $P(w|v)$ represents the probability of translating word v into word w. $P(w|v)$ is heuristically calculated as

$$P(w|v) = \frac{\mathrm{sim}(w,v)}{\sum_{w\in\mathcal{V}}\mathrm{sim}(w,v)}, \tag{5}$$

where $\mathrm{sim}(w,v)$ is the word similarity between w and v. In order to calculate $P(w|v)$, cosine distances between pre-trained word embeddings were recently utilized (Zuccon et al., 2015; Ganguly et al., 2015). Thus, the word similarity is calculated as

$$\mathrm{sim}(w,v) = \boldsymbol{w}^\top\boldsymbol{v}, \tag{6}$$

where $\boldsymbol{w}$ is the word embedding normalized to a unit length for w.

4 IR based on Hyperspherical QLMs

This paper proposes a novel probabilistic IR method based on hyperspherical QLM that leverages pre-trained word embeddings as random variables for directional probabilistic models. The pre-trained word embeddings can capture semantic similarity of words on a unit hypersphere, so we deal with normalized word embeddings to unit length. Given a query $Q = \{w_1, \cdots, w_T\}$, normalized word embeddings $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_T\}$ can be acquired using embedding models. IR based on hyperspherical QLMs ranks documents as

$$P(D|Q) \overset{\text{rank}}{\propto} \prod_{t=1}^{T} p(\boldsymbol{w}_t|\boldsymbol{\Lambda}_D), \tag{7}$$

where p means a probability density and $\boldsymbol{\Lambda}_D$ denotes a parameter of a hyperspherical QLM for D.

4.1 Formulation

The hyperspherical QLMs are formulated by a mixture of von Mises-Fisher distributions which are familiar to cosine distances (Banerjee et al., 2005; Sra, 2016). The von Mises-Fisher distribution defines a probability density over points on a unit hypersphere. A probability density based on the mixture of von Mises-Fisher distributions is formulated as

$$p(\boldsymbol{w}|\boldsymbol{\Lambda}) = \sum_{m=1}^{M} \alpha_m f(\boldsymbol{w}|\boldsymbol{\mu}_m), \tag{8}$$

$$f(\boldsymbol{w}|\boldsymbol{\mu}_m) = C_d(\kappa)\exp(\kappa \boldsymbol{w}^\top \boldsymbol{\mu}_m), \tag{9}$$

where M is the number of mixtures. The parameter $\boldsymbol{\Lambda}$ corresponds to $\{\alpha_m, \boldsymbol{\mu}_m\}$. α_m means a mixture weight, and $\boldsymbol{\mu}_m$ means a directional mean of the m-th von Mises-Fisher distribution. κ is a concentration parameter that is treated as a smoothing parameter. $C_d(\kappa)$ is a normalized parameter that depends on κ and the number of dimensions of word embeddings d. Note that $\boldsymbol{w}^\top \boldsymbol{\mu}_m$ is the cosine distance between $\boldsymbol{w}$ and $\boldsymbol{\mu}_m$.

4.2 ML and MAP Estimation for Mixture of von Mises-Fisher Distributions

ML estimation for a mixture of von Mises-Fisher distributions given normalized word embeddings $\boldsymbol{D} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{|D|}\}$ determines model parameters $\boldsymbol{\Lambda}_D^{\texttt{ML}}$ as

$$\boldsymbol{\Lambda}_D^{\texttt{ML}} = \underset{\boldsymbol{\Lambda}}{\operatorname{argmax}}\, P(\boldsymbol{D}|\boldsymbol{\Lambda}). \tag{10}$$

The ML estimation is based on the expectation maximization algorithm. ML estimated parameters $\boldsymbol{\Lambda}_D^{\texttt{ML}} = \{\hat{\alpha}_m\}, \{\hat{\boldsymbol{\mu}}_m\}$ are recursively calculated as

$$\hat{\alpha}_m = \frac{1}{|D|}\sum_{t=1}^{|D|} q(m|\boldsymbol{w}_t, \boldsymbol{\Lambda}), \tag{11}$$

$$\hat{\boldsymbol{r}}_m = \sum_{t=1}^{|D|} \boldsymbol{w}_t q(m|\boldsymbol{w}_t, \boldsymbol{\Lambda}), \tag{12}$$

$$\hat{\boldsymbol{\mu}}_m = \frac{\hat{\boldsymbol{r}}_m}{||\hat{\boldsymbol{r}}_m||}, \tag{13}$$

where $q(m|\boldsymbol{w}_t, \boldsymbol{\Lambda})$ is a load factor of the m-th distribution for the t-th word embedding.

MAP estimation for a mixture of von Mises-Fisher distributions given normalized word embeddings $\boldsymbol{D} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{|D|}\}$ determines model parameters $\boldsymbol{\Lambda}_D^{\texttt{ML}}$ as

$$\boldsymbol{\Lambda}_D^{\texttt{MAP}} = \underset{\boldsymbol{\Lambda}}{\operatorname{argmax}}\, P(\boldsymbol{D}|\boldsymbol{\Lambda})P(\boldsymbol{\Lambda}). \tag{14}$$

Given pre-trained parameters $\bar{\boldsymbol{\Lambda}} = \{\bar{\alpha}_m, \bar{\boldsymbol{r}}_m\}$, MAP-estimated parameters $\boldsymbol{\Lambda}_D^{\texttt{MAP}} = \{\hat{\alpha}_m, \hat{\boldsymbol{\mu}}_m\}$ are calculated as

$$\hat{\alpha}_m = \frac{\tau}{|D|+\tau}\bar{\alpha}_m + \frac{|D|}{|D|+\tau}\frac{1}{|D|}\sum_{t=1}^{|D|} q(m|\boldsymbol{w}_t, \bar{\boldsymbol{\Lambda}}), \tag{15}$$

$$\hat{\boldsymbol{r}}_m = \frac{\tau}{|D|+\tau}\bar{\boldsymbol{r}}_m + \frac{|D|}{|D|+\tau}\sum_{t=1}^{|D|} \boldsymbol{w}_t q(m|\boldsymbol{w}_t, \bar{\boldsymbol{\Lambda}}), \tag{16}$$

$$\hat{\boldsymbol{\mu}}_m = \frac{\hat{\boldsymbol{r}}_m}{||\hat{\boldsymbol{r}}_m||}, \tag{17}$$

where τ is a hyper parameter for adjusting smoothing.

For computation of load factors, both the soft-assignment rule and the hard-assignment rule can be used. Both computations are defined as

$$q_{\texttt{s}}(m|\boldsymbol{w}_t, \boldsymbol{\Lambda}) = \frac{\alpha_m f(\boldsymbol{w}_t|\boldsymbol{\mu}_m)}{\sum_{l=1}^{M} \alpha_l f(\boldsymbol{w}_t|\boldsymbol{\mu}_l)}, \tag{18}$$

$$q_{\texttt{h}}(m|\boldsymbol{w}_t, \boldsymbol{\Lambda}) = \begin{cases} 1 & m = \underset{l}{\operatorname{argmax}}\, \alpha_l f(\boldsymbol{w}_t|\boldsymbol{\mu}_l), \\ 0 & \text{otherwise}, \end{cases} \tag{19}$$

where $q_{\texttt{s}}$ is the load factor using the soft-assignment rule and $q_{\texttt{h}}$ is the load factor using the hard-assignment rule.

4.3 Training of Hyperspherical QLMs

In order to introduce a mixture of von Mises-Fisher distributions to hyperspherical QLMs, unified assumptions are essential for each document modeling because the hyperspherical QLMs are utilized for IR. Therefore, we introduce the following assumptions.

- The number of mixtures corresponds to vocabulary size.

  $$M = |\mathcal{V}|. \tag{20}$$

- The mean direction of each von Mises-Fisher distribution is fixed to normalized word embeddings of each word in the vocabulary.

  $$\hat{\boldsymbol{\mu}}_m = \boldsymbol{v}_m, \tag{21}$$

  where $v_m \in \mathcal{V}$.

- For computation of load factors in document modeling, a hard-assignment rule is used.

To summarize the above, the hyperspherical QLMs can be theoretically estimated as simple forms using ML or MAP estimation. In fact, mixture weights are estimated as ML or MAP estimated values in categorical QLMs. In ML estimation, $\hat{\alpha}_m$ estimated from document D is determined as

$$\hat{\alpha}_m = \frac{c(v_m, D)}{|D|}. \tag{22}$$

Thus, ML estimated generative probability of $\boldsymbol{w}$ in the hyperspherical QLMs is formulated as

$$p(\boldsymbol{w}|\boldsymbol{\Lambda}_D^{\texttt{ML}}) = \sum_{v \in \mathcal{V}} \frac{c(v, D)}{|D|} f(\boldsymbol{w}|\boldsymbol{v}), \tag{23}$$

where $\boldsymbol{v}$ is a normalized word embedding of v.

In MAP estimation, $\hat{\alpha}_m$ estimated from document D is determined as

$$\hat{\alpha}_m = \frac{c(v_m, D) + \tau \frac{c(v_m, C)}{|C|}}{|D| + \tau}, \tag{24}$$

where C is document collection in which all retrieved documents are included. Thus, MAP estimated generative probability of $\boldsymbol{w}$ in the hypershperical QLMs is formulated as

$$p(\boldsymbol{w}|\boldsymbol{\Lambda}_D^{\texttt{MAP}}) = \sum_{v \in \mathcal{V}} \frac{c(v, D) + \tau \frac{c(v, C)}{|C|}}{|D| + \tau} f(\boldsymbol{w}|\boldsymbol{v}). \tag{25}$$

4.4 Relationships

The hyperspherical QLMs can be interpreted as an extended form of categorical QLMs. Eq. (23) includes the ML estimated term presented in Eq. (2), and Eq. (25) includes the MAP estimated term presented in Eq. (3). In fact, hyperspherical QLMs can be converted into categorical QLMs by

$$\lim_{\kappa \to \infty} p(\boldsymbol{w}|\boldsymbol{\Theta}_D) = P(w|\boldsymbol{\Lambda}_D). \tag{26}$$

In addition, hyperspherical QLMs can be regarded as a theoretical form of translation QLMs. Eqs. (23) and (25) are similar to Eq. (4). In fact, hyperspherical QLMs are almost the same as translation QLMs by defining a word similarity as

$$\mathrm{sim}(w, v) = \exp(\kappa \boldsymbol{w}^\top \boldsymbol{v}). \tag{27}$$

While Eq. (6) is heuristically formulated, Eq. (27) is theoretically formulated as a log-linear form based on the directional probabilistic modeling.

5 Experiments

5.1 Setups

We performed an experiment on a document retrieval task, in which we used 20 news group datasets[1] for evaluation. The datasets were formally split into 11,314 training and 7,531 test articles. The training articles were used for collecting documents and the test articles were used for queries. Label information about news groups was only utilized for deciding whether a retrieved document is relevant to the query in evaluation. These setups are equivalent to the evaluation in Salakhutdinov and Hinton (2009); Larochelle and Lauly (2012). We removed common stop words, and the 5,000 most frequent words in the training articles were used for the vocabulary.

In order to utilize word embeddings, data sets in a one billion word language modeling benchmark[2] were prepared. We constructed CBOW and PV with distributed BOW (PV-DBOW). These pretrained embedding models were utilized for IR-methods. The dimension of the word and document embeddings was set to 200.

For evaluation, the following IR methods were used. **TFIDF** used cosine distance between two document vectors composed by word TF-IDF values. **CWV** used cosine distance between two

[1] http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz

[2] http://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark

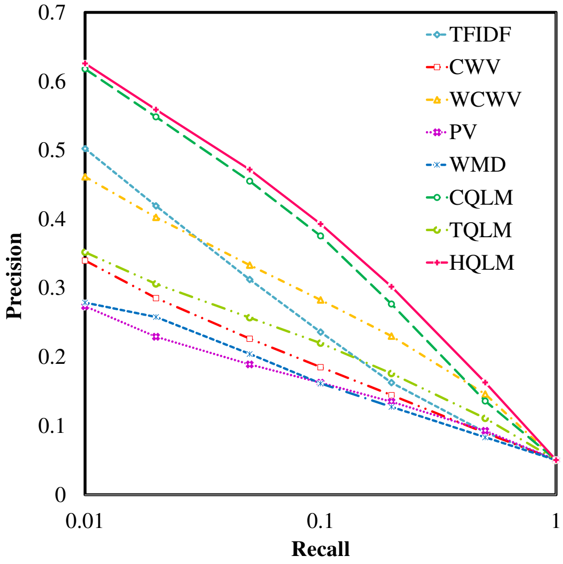

Figure 1: Precision-recall curves results.

document vectors composed by averaging word embeddings (Vulic and Moens, 2015). **WCWV** used cosine distance between two document vectors composed by adding word embeddings with IDF weights (Brokos et al., 2016). **PV** used cosine distance of two document vectors composed by PV-DBOW (Le and Mikolov, 2014). **WMD** used WMD using word embeddings (Kusner et al., 2015). **CQLM** used discrete QLMs estimated by Eq. (4) where τ was set to 2,000 (Zhai and Lafferty, 2001). **TQLM** used translation QLMs where word similarity was calculated by Eq. (6) using word embeddings, and Eq. (3) was used for calculating generative probability (Ganguly et al., 2015). **HQLM** used hyperspherical QLMs estimated by Eq. (11) using word embeddings, in which τ and κ were respectively set to 2,000 and 20.

5.2 Results

We assessed the performance of each IR-method with precision-recall curves, mean average precision (mAP), and precision at 10 (P@10). Figure 1 shows the precision-recall curve results and Table 1 shows the mAP and P@10 results. The results are averaged over all possible queries.

The results show that CQLM and HQLM clearly outperformed document distance metric-based IR methods with word or document embeddings. This confirms that QLM-based IR is a helpful approach for document retrieval. Although word or document embeddings trained from a lot of text sets can efficiently capture semantic information in continuous space, the document distance metrics using the embeddings were insufficient for document retrieval. In addition, we attained superior performance only introducing HQLM compared to CQLM while single use of TQLM did not perform well. In fact, previous work attained performance improvements by combining TQLM with CQLM. These results confirm that HQLM can effectively utilize word embeddings for document retrieval. Furthermore, we analyzed that relevant documents ranked low in CQLM were moved up by HQLM. This indicates HQLM can robustly calculate generative probabilities of words that were not included in a target query. We also verified that HQLM showed similar results to CQLM when κ was set to a large value ($\kappa = 500$). This confirms that Eq. (12), which insists HQLM can be represented as CQLM, is a proper theory.

Table 1: mAP and P@10 results.

IR methods	Embeddings	mAP	P@10
TFIDF		0.123	0.478
CWV	√	0.109	0.321
WCWV	√	0.165	0.438
PV	√	0.105	0.284
WMD	√	0.103	0.287
CQLM		0.182	0.586
TQLM	√	0.126	0.343
HQLM	√	**0.198**	**0.594**

6 Conclusions

In this paper, we proposed a word embedding-based probabilistic IR method based on hyperspherical QLMs that are modeled by a mixture of von Mises-Fisher distributions. We found the hyperspherical QLMs could theoretically utilize word embeddings for IR without introducing heuristic formulations. We found that the hyperspherical QLMs can be represented as an extended form of categorical QLMs and a theoretical form of translation QLMs. Our experiments on a document retrieval task showed hyperspherical QLM outperformed previous QLMs and document distance metrics with word or document embeddings. In the future, we will examine large scale document retrieval evaluation.

References

Jing Bai, Dawei Song, Peter Bruza, Jian-Yun Nie, and Guihong Cao. 2005. Query expansion using term relationships in language models for information retrieval. *In Proc. Conference on Information and Knowledge Management (CIKM)* pages 688–695.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von mises-fisher distribution. *Journal of Machine Learning Research* 6:1345–1382.

Nematollah Kayhan Batmanghelich, Ardavan Saeedi, Karthik R. Narasimhan, and Samuel J. Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)* pages 537–542.

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. *In Proc. Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)* pages 222–229.

Georgios-Ioannis Brokos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2016. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. *In Proc. Workshop on Biomedical Natural Language Processing (BioNLP)* pages 114–118.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)* pages 795–804.

Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth J.F. Jones. 2015. A word embedding based generalized language model for information retrieval. *In Proc. Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)* pages 795–798.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. *In Proc. International Conference on Machine Learning (ICML)* .

Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query expansion using word embeddings. *In Proc. Conference on Information and Knowledge Management (CIKM)* pages 1929–1932.

Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. *In Proc. Advances in Neural Infomation Processing Systems (NIPS)* pages 2708–2716.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *In Proc. International Conference on Machine Learning (ICML)* pages 1188–1196.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Ggeg Corrado, and Jeffrey Dean. 2013. Distributed representation of words and phrases and their compositionality. *In Proc. Advances in Neural Information Processing Systems (NIPS)* pages 3111–3119.

Bhaskar Mitra and Nick Craswell. 2017. Neural text embeddings for information retrieval. *In Proc. ACM International Conference on Web Search and Data Mining (WSDM)* pages 813–814.

Andriy Mnih and Koray Kavukcuglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. *In Proc. Advances in Neural Information Processing Systems (NIPS)* pages 2265–2273.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *In Proc. Empirical Methods in Natural Language Processing (EMNLP)* pages 1532–1543.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. *In Proc. Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)* pages 275–281.

Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Replicated softmax: an undirected topic model. *In Proc. Advances in Neural Information Processing Systems (NIPS)* pages 1607–1614.

Suvrit Sra. 2016. Directional statistics in machine learning: a brief review. *arXiv prerint arXiv:1605.00316* .

Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. *In Proc. Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)* pages 363–372.

Xing Wei and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. *In Proc. Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)* pages 178–185.

Hamed Zamani and W. Bruce Croft. 2016a. Embedding-based query language models. *In Proc. ACM International Conference on the Theory of Information Retrieval (ICTIR)* pages 147–156.

Hamed Zamani and W. Bruce Croft. 2016b. Estimating embedding vectors for queries. *In Proc. ACM International Conference on the Theory of Information Retrieval (ICTIR)* pages 123–132.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. *In Proc. Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)* pages 334–342.

Ye Zhang, Md Mustafizur Rahman, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, and Matthew Lease. 2016. Neural information retrieval: A literature review. *arXiv prerint arXiv:1611.06792* .

Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and evaluating neural word embeddings in information retrieval. *In Proc. Australasian Document Computing Symposium (ADCS)* 12:1–8.