Joint Learning from Labeled and Unlabeled Data for Information Retrieval

Bo Li, Ping Cheng, Le Jia
School of Computer Science
Central China Normal University
Wuhan, China
libo@mail.ccnu.edu.cn

Abstract

Recently, a significant number of studies have focused on neural information retrieval (IR) models. One category of works use unlabeled data to train general word embeddings based on term proximity. The general embeddings can be integrated into traditional IR models. The other category employs labeled data (e.g. click-through data) to train *end-to-end* neural IR models consisting of layers for target-specific representation learning. The latter idea accounts better for the IR task and is favored by recent research works, which is the one we will follow in this paper. We hypothesize that general semantics learned from unlabeled data can complement task-specific representation learned from labeled data of limited quality, and that a combination of the two is favorable. To this end, we propose a learning framework which can benefit from both labeled and more abundant unlabeled data for representation learning in the context of IR. Through a joint learning fashion in a single neural framework, the learned representation is optimized to minimize both the supervised loss on query-document matching and the unsupervised loss on text reconstruction. Standard retrieval experiments on TREC collections indicate that the joint learning methodology leads to significant better performance of retrieval over several strong baselines for IR.

1 Introduction

In recent years, the research community has noticed the great success of neural networks in computer vision (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012) and natural language processing (Mikolov et al., 2013) tasks. However, the potential of neural networks has not been fully investigated in the IR field. Although a significant number of studies (e.g. (Huang et al., 2013; Ganguly et al., 2015; Zheng and Callan, 2015; Guo et al., 2016; Zamani and Croft, 2016; Dehghani et al., 2017; Mitra et al., 2017)) try to apply neural networks in IR, there have been few studies reporting the performance that is comparable to state-of-the-art IR models. These approaches rely on the general idea that neural network can provide a low-dimensional and semantics-rich representation for both queries and documents. Such a representation can bridge lexical and semantic gaps in traditional IR models. Depending on if the embeddings are trained with discriminative information for IR tasks, existing works can be broadly divided into two categories (Zhang et al., 2016; Mitra and Craswell, 2017).

The first category of approaches extend traditional IR models to incorporate word embeddings that are trained on huge and unlabeled corpora with existing models such as *Word2vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014) in an unsupervised manner. These approaches (e.g. (Zheng and Callan, 2015; Nalisnick et al., 2016)) leverage semantic information captured by word embeddings in order to enhance traditional IR models. We note that such models trained without references to the retrieval task model term proximity and do not contain discriminative information adapted for IR (Zamani and Croft, 2017). The second category (e.g. (Huang et al., 2013; Guo et al., 2016)) tries to incorporate word embedding learning within neural models for IR, which reflects a more significant shift toward an *end-to-end* framework. These approaches treat word embeddings as layers in neural IR models, to be learned

Proceedings of the 27th International Conference on Computational Linguistics, pages 293–302
Santa Fe, New Mexico, USA, August 20-26, 2018.

along with all model parameters in a supervised manner. Most studies in the second category rely on click-through data for relevance judgment between queries and documents. Text representation learned with relevance information captures relevance rather than term proximity, which clearly accounts better for IR requirements (Zamani and Croft, 2017). However, supervised signals such as click-through data are often limited outside of large industrial research labs, probably due to user privacy concerns. It is thus not surprising to see that many authors following this methodology have industrial background (e.g. (Huang et al., 2013; Shen et al., 2014b; Nalisnick et al., 2016; Mitra et al., 2017)). In addition, Ye et al. (2015) point out that previous studies using click-through data make implicit but strong assumptions about clicked query-document pairs which are not necessarily met in practice.

Neural networks are hungry for data, a fact which also holds for neural IR tasks. One can find from above discussions that the second category of approaches suffer from the data spareness problem, although there have been recent attempts (Gupta et al., 2017; Dehghani et al., 2017) trying to pseudo label query-document pairs automatically with unsupervised retrieval models such as BM25. Using pseudo labels as relevance signals relieves data spareness in terms of quantity but not quality. The idea of using unsupervised learning to complement supervision has been practiced successfully in computer vision (Yang et al., 2013) and natural language processing (Rasmus et al., 2015) tasks. In such a background, we hypothesize that semantics learned from unlabeled data can complement task-specific representation learned from pseudo-labeled data of limited quality, and a combination of the two is favorable in IR. To the best of our knowledge, such a combination has never been investigated in neural IR models.

In this paper, we propose a learning framework which can benefit from both labeled and more abundant unlabeled data for representation learning in IR. Through joint learning in a single neural network, the learned representation can account for task-specific characteristics via supervised loss optimization on query-document matching, as well as preserving general semantics via unsupervised loss optimization on text reconstruction. We demonstrate by experiments that the joint learning model leads to significantly better performance over state-of-the-art IR models.

2 Related work

Representation learning approaches based on neural networks have gained in prominence in recent years due to their extreme efficiency. They motivate the emerging research field of Neural IR. Neural approaches have attracted increasing interests of the IR community in very recent years. Apart from learning to rank approaches that train their models over a set of hand-crafted features (Liu, 2009), neural IR models typically accept the raw text of queries and documents as input. The dense representations of words or texts can then be learned with or without reference to retrieval tasks, respectively corresponding to the two categories of methods summarized in section 1.

Unsupervised approaches learn general text representation without query and document interaction information. Embeddings pre-trained on unlabeled text with tools such as *Word2vec* (Mikolov et al., 2013) and *Glove* (Pennington et al., 2014) have been used to extend traditional IR models. Ganguly et al. (2015) develop a generalized language model with query-likelihood language modeling for integrating word embeddings as additional smoothing. Zheng and Callan (2015) represent term and query as vectors in the same latent space based on word embeddings so as to learn a model to reweight terms. Nalisnick et al. (2016) retain both input and output embeddings of *Word2vec* and map query words into the input space and document words into the output space. Zamani and Croft (2016) propose to use word embeddings to incorporate and weight terms not present in the query, acting as smoothing and query expansion. There are also studies developing their own embedding learning algorithms instead of using standard tools for embedding learning. For instance, Salakhutdinov and Hinton (2009) propose a deep auto-encoder model to generate a condensed binary vector representation of documents. Clinchant and Perronnin (2013) use latent semantic indexing to induce word embeddings for IR. Vulić and Moens (2015) propose to learn from document-aligned comparable corpora the embeddings that can be used for both monolingual IR and cross-lingual IR.

Supervised approaches use query-document relevance information to learn the representation that is optimized *end-to-end* for the task at hand. With click-through data, Huang et al. (2013) develop DSSM, a feed forward neural network with a word hashing phrase as the first layer to predict the click probability given a query string and a document title. DSSM is extended in (Shen et al., 2014a; Shen et al., 2014b) by incorporating convolutional neural network and max-pooling layers to extract the most salient local features. Since the DSSM related methods make implicit but strong assumptions about clicked data, Ye et al. (2015) try to relax the assumptions in their model. Guo et al. (2016) develop the DRMM model that takes the histogram-based features representing interactions between queries and documents as input into neural networks. DRMM is one of the first neural IR models to show improvement over traditional IR models. Mitra et al. (2017) aim to simultaneously learn local and distributional representation to capture both lexical matching and semantic matching in IR. Following the discussion in section 1, we note that click-through data are not always available in massive amount outside of industrial labs. More recent works propose to use unsupervised IR models to pseudo label query-document pairs that provide weak supervision for representation learning. Dehghani et al. (2017) use BM25 to obtain relevant documents for a large set of AOL queries (Pass et al., 2006) which are then used as weakly supervised signals for joint embedding and ranking model training. Zamani and Croft (2017) employ similar supervision signals as (Dehghani et al., 2017) to train an embedding network similar to *Word2vec* and use the obtained embeddings for query expansion and query classification. Gupta et al. (2017) develop a cross-lingual IR model based on weak supervision. Luo et al. (2017) propose to train deep ranking models with weak relevance labels generated by click model based on click behavior of real users.

We can conclude from above discussions that supervised approaches account better for task-specific features and are superior in IR. They rely on relevance information between query-document pairs of which the quality is relatively low in practice. In this paper, we follow successful practice in CV and NLP tasks and hypothesize that general and rich semantics learned from unlabeled data can complement task-specific representation learned from labeled data of limited quality. We will propose in section 3 a learning framework which can simultaneously learn from labeled and more abundant unlabeled data in the context of IR. By the way, we note that the joint learning framework resembles those studies (e.g. (Liu et al., 2015)) which couple IR with another supervised learning task. Our framework differs from those studies in that we do not require additional data that are labeled for another supervised learning task.

3 Joint learning framework for IR

In this section, we will develop a joint framework to learn low-dimensional representation of queries and documents from both labeled and unlabeled data.

3.1 Learning framework

The joint learning framework is illustrated in figure 1. It consists of three crucial components:

- **An encoding network.** It embeds the raw input into low-dimensional representations that are designed to capture target-specific characteristics of IR.

- **A decoding network.** It tries to reconstruct the input so as to benefit from unlabeled data.

- **A pairwise ranking model.** It makes use of supervision signals from labeled query-document pairs to perform document ranking.

On top of the network structure, we perform joint optimization of both supervised loss and unsupervised loss. The unsupervised learning process uses all the text collection (e.g. queries and documents) for learning rich and general semantics. The supervised learning process learns, from labeled query-document pairs, discriminative representations adapted for IR. The joint training fashion makes two learning processes complement each other via co-tuning the shared hidden layers in the encoding networks to help the representation generalize better in the IR task.

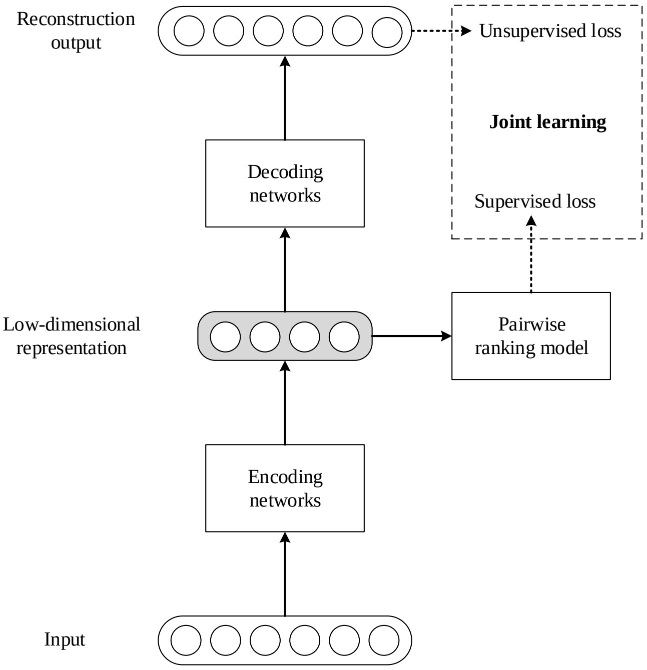

Figure 1: The joint learning framework with labeled and unlabeled data. It consists of an encoding network, a decoding network and a pairwise ranking model. We impose an unsupervised loss and a supervised loss respectively on the reconstruction output and the pairwise ranking output, which is learned in a joint fashion in this paper. The low-dimensional representation is the one our model aims to learn.

3.2 Unsupervised learning

The unsupervised part learns the low-dimensional representation of text via an autoencoder style, which uses all the available text data. Following previous studies on text autoencoder (Chen and Zaki, 2017), we opt for the simple feed-forward neural network architecture for both the encoding and decoding parts in figure 1. For each layer of the encoding/decoding networks, we use Rectified Linear Unit (ReLU) as the activation function, a function recommended by many works in deep learning (LeCun et al., 2015). In the feed-forward step, each layer $l(l \geq 1)$ is a fully-connected layer and its activation potential $\boldsymbol{z}_l$ is given by:

$$\boldsymbol{z}_l = \max(0, \boldsymbol{W}_l \boldsymbol{z}_{l-1} + \boldsymbol{b}_l)$$

where $\boldsymbol{W}_l$ is the weight matrix at layer l and $\boldsymbol{b}_l$ is the corresponding bias.

The input layer (corresponding to $l = 0$) maps the input text into fixed-length vector. There have been two methodologies we can employ to represent the input text: one is the one-hot representation (Gupta et al., 2017) and its variants (Zhai and Zhang, 2016); the other one is the dense and semantically rich representations (He et al., 2017). Empirical results do not indicate that one is always better than the other and we will make use of the former one in this paper. Given the set of text T, we follow previous studies such as (Zhai and Zhang, 2016; Chen and Zaki, 2017) and represent each input text t in T as log-normalized word count vector $\boldsymbol{x} \in R^{|V|}$ where $|V|$ is the size of the vocabulary V. Each dimension of the input vector $\boldsymbol{x}$ is represented by:

$$\boldsymbol{x}_i = \frac{\log[1 + tf(i)]}{\max_{i \in V} \log[1 + tf(i)]}, \text{ for } i \in V$$

where $tf(i)$ is the term frequency of the i-th word in the vocabulary. Since the unsupervised learning part of the framework is modeled as an autoencoder, we want the unsupervised output $\boldsymbol{x}'$ to resemble the input $\boldsymbol{x}$, leading to the binary cross-entropy loss function l_u on t that can be defined as:

$$l_u(t) = -\sum_{i \in V} [\boldsymbol{x}_i \log(\boldsymbol{x}'_i) + (1 - \boldsymbol{x}_i) \log(1 - \boldsymbol{x}'_i)] \tag{1}$$

3.3 Supervised learning

The document ranking problem can not be modeled with the standard classification or regression framework. Following the methodology in learning to rank (Liu, 2009), we model document ranking in the *pairwise* style where the relevance information is in the form of preferences between pairs of documents with respect to individual queries. In addition, we follow previous studies (Gupta et al., 2017) and make use of well-performing unsupervised retrieval models (e.g. BM25) to pseudo-label query and document pairs so as to obtain the relevance information. More details will be given in section 4.1.

From figure 1 one can note that the hidden layers in the encoding networks are shared by unsupervised and supervised learning, and one can refer to the unsupervised learning part for details of the layers in the encoding networks. The supervised model, on top of the top-level representation layer (i.e. low-dimensional representation), tries to learn a model that, given the query q, assigns a larger score to document d_1 than document d_2 if the ground truth is that d_1 matches to q better. The supervised model is implemented as a pairwise ranking model in figure 1, which is again a feed forward neural networks. Inspired by such studies as (Yih et al., 2011), we can derive the probability $P'(d_1 \succ_q d_2)$ that d_1 is ranked higher than d_2 with respect to the query q via a logistic function:

$$P'(d_1 \succ_q d_2) = \frac{1}{1 + e^{-\sigma[\text{score}(q,d_1) - \text{score}(q,d_2)]}}$$

where the *score* function is computed with the pairwise ranking model, and the parameter σ is used to determine the shape of the sigmoid. The supervised training objective l_s on a triplet of query-document pair (q, d_1, d_2) can then be defined as the cross entropy loss, which is:

$$\begin{aligned} l_s(q, d_1, d_2) = &- P(d_1 \succ_q d_2) \log P'(d_1 \succ_q d_2) \\ &- [1 - P(d_1 \succ_q d_2)] \log[1 - P'(d_1 \succ_q d_2)] \end{aligned} \tag{2}$$

where $P(d_1 \succ_q d_2)$ is the actual probability that d_1 is ranked higher than d_2 according to annotations (i.e. pseudo-labels of query-document pairs). The actual probability in this paper is estimated in a similar way as in (Dehghani et al., 2017), which is:

$$P(d_1 \succ_q d_2) = \frac{1}{1 + e^{-\sigma[\text{S}(q,d_1) - \text{S}(q,d_2)]}}$$

where s denotes the relevance scores obtained from training instances. In the training process, the positive sample d_1 for the query q can be chosen as the most relevant documents according to annotated relevance scores. The negative sample d_2 is selected randomly from the document collection.

3.4 Joint learning with regularization

Combining the unsupervised loss l_u in equation 1 on all text data, the supervised loss l_s in equation 2 on all labeled query-document pairs, and the $L2$ norm regularization for weight matrices, one finally arrives at the objective function for the joint learning model, which is:

$$\begin{aligned} L(T, DS) = &\frac{\alpha}{|T|} \sum_{t \in T} l_u(t) + \frac{\beta}{|QD|} \sum_{(q,d_1,d_2) \in QD} l_s(q, d_1, d_2) \\ &+ \sum_{l \in LY} \|W_l\|_F^2 \end{aligned} \tag{3}$$

where T and $|T|$ denote the set of text data and its size, QD and $|QD|$ denote the set of labeled query-document pairs and its size, LY stands for all the hidden and output layers of the framework in figure 1, and W_l is the weight matrix of the layer l in the network. The hyper-parameters α, β control the importance of the unsupervised loss and the supervised loss. The joint loss function $L(T, DS)$ can be optimized in the gradient-based way, and we use the Adam algorithm (Kingma and Ba, 2015) to compute the gradients.

4 Experiments and results

In this section, we conduct IR experiments to demonstrate the effectiveness of our proposed model.

4.1 Data sets

The IR experiments are carried out against standard TREC collections consisting of one Robust track and one Web track, which represent different sizes and genres of heterogeneous text collections. These collections have been broadly used in recent studies (Zheng and Callan, 2015; Guo et al., 2016; Dehghani et al., 2017). The details of these collections and corresponding queries are given in table 1. The Robust dataset is used in the standard form without change. The ClueWeb-09-Cat-B collection (or *ClueWeb* for short) is filtered to the set of documents with spam scores in the 60-th percentile with Waterloo Fusion spam scores[1]. For all TREC queries, we only make use of the title fields for retrieval.

Table 1: IR collection statistics (M = million, B=Billion).

Collections	Doc count	Word count	TREC topics
Robust04	0.5M	252M	301-450, 601-700
ClueWeb	34.0M	26.1B	1-200

In order to build the labeled query-document pairs for supervised learning, we choose to use the more general methodology in (Gupta et al., 2017) instead of the one in (Dehghani et al., 2017) to relieve from data (i.e. AOL queries) only available from industrial labs. We fetch a set of news titles from the China Daily website[2] and use these titles as training queries to produce annotated query-document pairs. We use these training queries to retrieve the document collection with BM25. We make sure that no training queries appear in the evaluation query set in table 1. For each training query, we take the top 500 retrieved documents as positive samples. The negative samples are picked randomly from the document collection. There are other strategies for choosing negative samples (Wieting et al., 2015), which is out of the scope of this paper. For unsupervised learning, we make use of training queries and evaluation document sets listed in table 1, as well as the Wikipedia articles[3] as the external resource.

4.2 Experimental setup

We set the hyper-parameters of our model by following similar tasks such as (Dehghani et al., 2017). The size and number of hidden layers are respectively selected from $\{64, 128, 256, 512, 1024\}$ and $\{1, 2, 3, 4\}$. The values of α, β in equation 3 are chosen from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. We select the initial learning rate from $\{10^{-3}, 10^{-4}, 5 * 10^{-4}, 10^{-5}, 5 * 10^{-5}\}$. The batch size for learning is selected from $\{64, 128, 256, 512\}$. These model hyper-parameters are tuned on the validation set (20% of the training queries used for validation).

For IR evaluation, we make use of mean average precision (MAP) of top-ranked 1000 documents , precision at rank 20 (P20), and normalized discounted cumulative gain at rank 20 (nDCG20). Statistically significant differences between various models are determined using the two-tailed paired t-test with $p < 0.05$.

We compare the retrieval performance of our joint learning retrieval model with two categories of IR models: classic IR models showing state-of-the-art performance, and the recent neural ranking models for IR. Since our model is representation-focused rather than interaction-focused, we do not plan to compare our model with those based on relevance matching (Guo et al., 2016) in this paper. More importantly, since our model learns from weakly supervised signals by BM25, we are more interested in the comparisons to BM25 and similar models using weakly supervised signals, an experimental strategy also employed in (Dehghani et al., 2017). Under such considerations, we perform experiments with the following baselines:

[1] https://plg.uwaterloo.ca/˜gvcormac/clueweb09spam

[2] http://www.chinadaily.com.cn

[3] The wikipedia dump on September 1, 2017 can be obtained from https://dumps.wikimedia.org

Table 2: Retrieval performance of all models on TREC collections. Significant improvement or degradation at the level 0.05 with respect to *BM25* is indicated as (+/-). The other significance comparisons are given in the text.

	Robust04			**ClueWeb**		
	MAP	P20	nDCG20	MAP	P20	nDCG20
BM25	0.248	0.351	0.406	0.091	0.237	0.190
QL	0.245	0.352	0.404	0.092	0.239	0.193
DSSM	0.088^{-}	0.163^{-}	0.184^{-}	0.037^{-}	0.126^{-}	0.104^{-}
NRMS	0.275^{+}	0.378^{+}	0.441^{+}	0.127^{+}	0.302^{+}	0.236^{+}
Our Model	$\mathbf{0.287}^{+}$	$\mathbf{0.391}^{+}$	$\mathbf{0.450}^{+}$	$\mathbf{0.136}^{+}$	$\mathbf{0.317}^{+}$	$\mathbf{0.251}^{+}$

- *Classic models:* The probabilistic *BM25* model and query likelihood (*QL*) model based on Dirichlet smoothing are highly efficient IR models.

- *DSSM:* It is a representative deep matching model proposed in (Huang et al., 2013), which is a representation-focused model. The model is framed as a feed forward neural network with a word hashing layer.

- *NRMS:* It is a weakly-supervised neural IR model learned with automatically annotated query-document pairs (Dehghani et al., 2017). NRMS shows significant improvement over traditional IR models.

4.3 Results and analysis

Comparisons to classic models. We use here the recommended settings of the baseline models according to their original papers. Table 2 reports the experimental results on TREC datasets for our model and all the baseline models. One can find from the results that classic IR models BM25 and QL perform similarly on the two collections, a conclusion that is coincident with previous findings. Since BM25 is the model we employ to produce pseudo labels for supervised learning, we will not compare neural models with QL in the following discussions. The neural IR model DSSM performs significantly worse than the traditional BM25 model, due to its unsuitability for relevance matching and for handling the diverse matching requirements in long documents (Guo et al., 2016). NRMS is a neural ranking model learned from automatically labeled data, which resembles our model. NRMS shows all the significant improvements over BM25. Our model proposed in this paper, by jointly learning from the labeled and unlabeled data, achieves the best overall performance. Our model always significantly outperforms BM25 by a large margin.

Comparisons to neural models. We further compare our model with the neural IR models DSSM and NRMS. We find that our model performs better than DSSM and NRMS on all collections. Our model significantly outperforms DSSM in all the cases considered above. Our model significantly outperforms NRMS with only one exception that is not significant on Robust04 with nDCG20. By the way, we find that NRMS is also always significantly better than DSSM on all collections. The experimental conclusion is that our model is always significantly better than traditional IR models and mostly outperforms neural IR models considered above. Furthermore, we find that using unlabeled data for training in neural IR models is useful, since it leads to significant improvement over the neural models only using labeled data.

Impact of unsupervised learning. It has been confirmed above that our model shows the best performance overall. However, it is not clear how much unsupervised learning contributes to the retrieval performance. We thus compare representations learned in a different setting without the help of unsupervised loss, which amounts to removing the unsupervised loss l_u from equation 3. We perform IR experiments with the new model over data sets in table 1 and list results in table 3. From the results one can find that the performance of the model without unsupervised loss decreases from the joint model with significance in all the cases considered. It indicates that it is beneficial to combine unsupervised

Table 3: Retrieval performance of the model without unsupervised loss. Significant degradation at the level 0.05 with respect to our original model is indicated as -.

	Robust04			**ClueWeb**		
	MAP	P20	nDCG20	MAP	P20	nDCG20
Original Model	0.287	0.391	0.450	0.136	0.317	0.251
Without unsupervised loss	0.262^-	0.356^-	0.413^-	0.114^-	0.298^-	0.231^-

learning with supervised learning in neural IR. Empirical results in this part support our claim in this paper that learning from unlabeled data complements knowledge learned from labeled data in neural IR.

5 Conclusions

In this paper, we propose a neural IR model which jointly learns from labeled and unlabeled data to benefit from both the rich and general semantics in unlabeled data and target-specific features in labeled data. As far as we can tell, it is the first time such a combination is investigated in neural IR. Experiments on TREC collections show that our model, without any human annotation, is significantly better than traditional IR models and recently proposed models based on neural networks. Experiments also show that using unsupervised learning to complement supervised learning with weak supervision is important in IR. A future direction to follow would be to use more expressive architectures such as LSTM to replace feed-forward networks used in this paper.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported by the Fundamental Research Funds for Central Universities of CCNU (No. CCNU15A05062).

References

Yu Chen and Mohammed J. Zaki. 2017. Kate: K-competitive autoencoder for text. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, pages 85–94.

Stephane Clinchant and Florent Perronnin. 2013. Aggregating continuous word embeddings for information retrieval. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 100–109.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 65–74.

Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth J.F. Jones. 2015. 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*, SIGIR, pages 795–798.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, CIKM, pages 55–64.

Parth Gupta, Rafael E. Banchs, and Paolo Rosso. 2017. Continuous space models for clir. *Information Processing and Management*, 53(2):359 – 370.

L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen. 2017. Unsupervised cross-modal retrieval through adversarial learning. In *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo*, ICME, pages 1153–1158.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM, pages 2333–2338.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS, pages 1097–1105.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521:436–444.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, pages 912–921.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Cheng Luo, Yukun Zheng, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Training deep ranking model with weak relevance labels. In *Proceedings of the Australasian Database Conference*, pages 205–216.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS, pages 3111–3119.

Bhaskar Mitra and Nick Craswell. 2017. Neural models for information retrieval. *CoRR*, abs/1705.01509.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW, pages 1291–1299.

Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. 2016. Improving document ranking with dual word embeddings. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW, pages 83–84.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, InfoScale.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1532–1543.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS, pages 3546–3554.

Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014a. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, CIKM, pages 101–110.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014b. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW, pages 373–374.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 363–372.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Y. Yang, G. Shu, and M. Shah. 2013. Semi-supervised learning of feature hierarchies for object detection in a video. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pages 1650–1657.

X. Ye, Z. Qi, and D. Massey. 2015. Learning relevance from click data via neural network based similarity models. In *Proceedings of the 2015 IEEE International Conference on Big Data*, pages 801–806.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL, pages 247–256.

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

Hamed Zamani and W. Bruce Croft. 2017. Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 505–514.

Shuangfei Zhai and Zhongfei Mark Zhang. 2016. Semisupervised autoencoder for sentiment analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI, pages 1394–1400.

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. *CoRR*, abs/1611.06792.

Guoqing Zheng and Jamie Callan. 2015. Learning to reweight terms with distributed representations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 575–584.