Information Processing and Management 57 (2020) 102067

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A Deep Look into neural ranking models for information retrieval

Jiafeng Guo[*,a,b], Yixing Fan[a,b], Liang Pang[a,b], Liu Yang[c], Qingyao Ai[c], Hamed Zamani[c], Chen Wu[a,b], W. Bruce Croft[c], Xueqi Cheng[a,b]

[a] University of Chinese Academy of Sciences, Beijing, China
[b] CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[c] Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, MA, USA

ARTICLE INFO

Keywords:
Neural ranking model
Information retrieval
Survey

MSC:
00-01
99-00

ABSTRACT

Ranking models lie at the heart of research on information retrieval (IR). During the past decades, different techniques have been proposed for constructing ranking models, from traditional heuristic methods, probabilistic methods, to modern machine learning methods. Recently, with the advance of deep learning technology, we have witnessed a growing body of work in applying shallow or deep neural networks to the ranking problem in IR, referred to as neural ranking models in this paper. The power of neural ranking models lies in the ability to learn from the raw text inputs for the ranking problem to avoid many limitations of hand-crafted features. Neural networks have sufficient capacity to model complicated tasks, which is needed to handle the complexity of relevance estimation in ranking. Since there have been a large variety of neural ranking models proposed, we believe it is the right time to summarize the current status, learn from existing methodologies, and gain some insights for future development. In contrast to existing reviews, in this survey, we will take a deep look into the neural ranking models from different dimensions to analyze their underlying assumptions, major design principles, and learning strategies. We compare these models through benchmark tasks to obtain a comprehensive empirical understanding of the existing techniques. We will also discuss what is missing in the current literature and what are the promising and desired future directions.

1. Introduction

Information retrieval is a core task in many real-world applications, such as digital libraries, expert finding, Web search, and so on. Essentially, IR is the activity of obtaining some information resources relevant to an information need from within large collections. As there might be a variety of relevant resources, the returned results are typically ranked with respect to some relevance notion. This ranking of results is a key difference of IR from other problems. Therefore, research on ranking models has always been at the heart of IR.

Ma ny different ranking models have been proposed over the past decades, including vector space models (Salton, Wong, & Yang, 1975), probabilistic models (Robertson & Jones, 1976), and learning to rank (LTR) models (Li, 2011; Liu, 2009). Existing techniques, especially the LTR models, have already achieved great success in many IR applications, e.g., modern Web search engines like

* Corresponding author at: CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.
E-mail address: guojiafeng@ict.ac.cn (J. Guo).

https://doi.org/10.1016/j.ipm.2019.102067
Received 9 March 2019; Received in revised form 8 June 2019; Accepted 20 June 2019
Available online 09 July 2019
0306-4573/

Google[1] or Bing[2]. There is still, however, much room for improvement in the effectiveness of these techniques for more complex retrieval tasks.

In recent years, deep neural networks have led to exciting breakthroughs in speech recognition (Hinton et al., 2012), computer vision (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015), and natural language processing (NLP) (Bahdanau, Cho, & Bengio, 2014; Goldberg, 2017). These models have been shown to be effective at learning abstract representations from the raw input, and have sufficient model capacity to tackle difficult learning problems. Both of these are desirable properties for ranking models in IR. On one hand, most existing LTR models rely on hand-crafted features, which are usually time-consuming to design and often over-specific in definition. It would be of great value if ranking models could learn the useful ranking features automatically. On the other hand, relevance, as a key notion in IR, is often vague in definition and difficult to estimate since relevance judgments are based on a complicated human cognitive process. Neural models with sufficient model capacity have more potential for learning such complicated tasks than traditional shallow models. Due to these potential benefits and along with the expectation that similar successes with deep learning could be achieved in IR (Craswell, Croft, Guo, Mitra, & de Rijke, 2017a), we have witnessed substantial growth of work in applying neural networks for constructing ranking models in both academia and industry in recent years. Note that in this survey, we focus on neural ranking models for textual retrieval, which is central to IR, but not the only mode that neural models can be used for (Brenner, Zhao, Kutiyanawala, & Yan, 2018; Wan et al., 2014).

Per haps the first successful model of this type is the Deep Structured Semantic Model (DSSM) (Huang et al., 2013) introduced in 2013, which is a neural ranking model that directly tackles the ad-hoc retrieval task. In the same year, Lu and Li (2013) proposed DeepMatch, which is a deep matching method applied to the Community-based Question Answering (CQA) and micro-blog matching tasks. Note that at the same time or even before this work, there were a number of studies focused on learning low-dimensional representations of texts with neural models (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013b; Salakhutdinov & Hinton, 2009) and using them either within traditional IR models or with some new similarity metrics for ranking tasks. However, we would like to refer to those methods as representation learning models rather than neural ranking models, since they did not directly construct the ranking function with neural networks. Later, between 2014 and 2015, work on neural ranking models began to grow, such as new variants of DSSM (Huang et al., 2013), ARC I and ARC II (Hu, Lu, Li, & Chen, 2014), MatchPyramid (Pang et al., 2016b), and so on. Most of this research focused on short text ranking tasks, such as TREC QA tracks and Microblog tracks (Severyn & Moschitti, 2015). Since 2016, the study of neural ranking models has bloomed, with significant work volume, deeper and more rigorous discussions, and much wider applications (Onal et al., 2018). For example, researchers began to discuss the practical effectiveness of neural ranking models on different ranking tasks (Cohen, Ai, & Croft, 2016; Guo, Fan, Ai, & Croft, 2016). Neural ranking models have been applied to ad-hoc retrieval (Hui, Yates, Berberich, & de Melo, 2017a; Mitra, Diaz, & Craswell, 2017), community-based QA (Qiu & Huang, 2015), conversational search (Yan, Song, & Wu, 2016a), and so on. Researchers began to go beyond the architecture of neural ranking models, paying attention to new training paradigms of neural ranking models (Dehghani, Zamani, Severyn, Kamps, & Croft, 2017b), alternate indexing schemes for neural representations (Zamani, Dehghani, Croft, Learned-Miller, & Kamps, 2018b), integration of external knowledge (Xiong, Callan, & Liu, 2017a; Yang et al., 2018), and other novel uses of neural approaches for IR tasks (Fan et al., 2017a; Tang & Yang, 2018).

Up to now, we have seen exciting progress on neural ranking models. In academia, several neural ranking models learned from scratch can already outperform state-of-the-art LTR models with tens of hand-crafted features (Fan et al., 2018; Pang et al., 2017). Workshops and tutorials on this topic have attracted extensive interest in the IR community (Craswell et al., 2017a; Craswell, Croft, de Rijke, Guo, & Mitra, 2017b). Standard benchmark datasets (Nguyen et al., 2016b; Yang, Yih, & Meek, 2015), evaluation tasks (Dietz, Verma, Radlinski, & Craswell, 2017), and open-source toolkits (Fan et al., 2017b) have been created to facilitate research and rigorous comparison. Meanwhile, in industry, we have also seen models such as DSSM put into a wide range of practical usage in the enterprise (He, Gao, & Deng, 2014). Neural ranking models already generate the most important features for modern search engines. However, beyond these exciting results, there is still a long way to go for neural ranking models: (1) Neural ranking models have not had the level of breakthroughs achieved by neural methods in speech recognition or computer vision; (2) There is little understanding and few guidelines on the design principles of neural ranking models; (3) We have not identified the special capabilities of neural ranking models that go beyond traditional IR models. Therefore, it is the right moment to take a look back, summarize the current status, and gain some insights for future development.

There have been some related surveys on neural approaches to IR (neural IR for short). For example, Onal et al. (2018) reviewed the current landscape of neural IR research, paying attention to the application of neural methods to different IR tasks. Mitra and Craswell (2017) gave an introduction to neural information retrieval. In their booklet, they talked about fundamentals of text retrieval, and briefly reviewed IR methods employing pre-trained embeddings and neural networks. In contrast to this work, this survey does not try to cover every aspect of neural IR, but will focus on and take a deep look into ranking models with deep neural networks. Specifically, we formulate the existing neural ranking models under a unified framework, and review them from different dimensions to understand their underlying assumptions, major design principles, and learning strategies. We also compare representative neural ranking models through benchmark tasks to obtain a comprehensive empirical understanding. We hope these discussions will help researchers in neural IR learn from previous successes and failures, so that they can develop better neural ranking models in the future. In addition to the model discussion, we also introduce some trending topics in neural IR, including indexing schema, knowledge integration, visualized learning, contextual learning and model explanation. Some of these topics are important but have

[1] http://google.com
[2] http://bing.com

not been well addressed in this field, while others are very promising directions for future research.

In the following, we will first introduce some typical textual IR tasks addressed by neural ranking models in Section 2. We then provide a unified formulation of neural ranking models in Section 3. From Sections 4–6 we review the existing models with regard to different dimensions as well as making empirical comparisons between them. We discuss trending topics in Section 7 and conclude the paper in Section 8.

2. Major applications of neural ranking models

In this section, we describe several major textual IR applications where neural ranking models have been adopted and studied in the literature, including ad-hoc retrieval, question answering, community question answering, and automatic conversation. There are other applications where neural ranking models have been or could be applied, e.g., product search (Brenner et al., 2018), sponsored search (Grbovic, Djuric, Radosavljevic, Silvestri, & Bhamidipati, 2015), and so on. However, due to page limitations, we will not include these tasks in this survey.

2.1. Ad-hoc retrieval

Ad-hoc retrieval is a classic retrieval task in which the user specifies his/her information need through a query which initiates a search (executed by the information system) for documents that are likely to be relevant to the user. The term *ad-hoc* refers to the scenario where documents in the collection remain relatively static while new queries are submitted to the system continually (Baeza-Yates, Ribeiro et al., 2011). The retrieved documents are typically returned as a ranking list through a ranking model where those at the top of the ranking are more likely to be relevant.

There has been a long research history on ad-hoc retrieval, with several well recognized characteristics and challenges associated with the task. A major characteristic of ad-hoc retrieval is the heterogeneity of the query and the documents. The query comes from a search user with potentially unclear intent and is usually very short, ranging from a few words to a few sentences (Mitra & Craswell, 2017). The documents are typically from a different set of authors and have longer text length, ranging from multiple sentences to many paragraphs. Such heterogeneity leads to the critical vocabulary mismatch problem (Furnas, Landauer, Gomez, & Dumais, 1987; Zhao & Callan, 2010). Semantic matching, meaning matching words and phrases with similar meanings, could alleviate the problem, but exact matching is indispensable especially with rare terms (Guo et al., 2016). Such heterogeneity also leads to diverse relevance patterns. Different hypotheses, e.g. verbosity hypothesis and scope hypothesis (Robertson & Walker, 1994), have been proposed considering the matching of a short query against a long document. The *relevance* notion in ad-hoc retrieval is inherently vague in definition and highly user dependent, making relevance assessment a very challenging problem.

For the evaluation of different neural ranking models on the ad-hoc retrieval task, a large variety of TREC collections have been used. Specifically, retrieval experiments have been conducted over neural ranking models based on TREC collections such as Robust (Guo et al., 2016; Pang et al., 2016b), ClueWeb (Guo et al., 2016), GOV2 (Fan et al., 2018; Pang et al., 2017) and Microblog (Pang et al., 2017), as well as logs such as the AOL log (Dehghani et al., 2017b) and the Bing Search log (Huang et al., 2013; Mitra et al., 2017; Palangi et al., 2016; Shen, He, Gao, Deng, & Mesnil, 2014). Recently, a new large scale dataset has been released, called the NTCIR WWW Task (Zheng et al., 2018), which is suitable for experiments on neural ranking models.

2.2. Question answering

Question-answering (QA) attempts to automatically answer questions posed by users in natural languages based on some information resources. The questions could be from a closed or open domain (Mollá & Vicedo, 2007), while the information resources could vary from structured data (e.g., knowledge base) to unstructured data (e.g., documents or Web pages) (Moschitti et al., 2016). There have been a variety of task formats for QA, including multiple-choice selection (Richardson, 2013), answer passage/sentence retrieval (Voorhees & Tice, 2000; Yang et al., 2015), answer span locating (Rajpurkar, Zhang, Lopyrev, & Liang, 2016), and answer synthesizing from multiple sources (Simon, Gao, Craswell, Deng). However, some of the task formats are usually not treated as an IR problem. For example, multiple-choice selection is typically formulated as a classification problem while answer span locating is usually studied under the machine reading comprehension topic. In this survey, therefore, we focus on answer passage/sentence retrieval as it can be formulated as a typical IR problem and addressed by neural ranking models. Hereafter, we will refer to this specific task as QA for simplicity.

Compared with ad-hoc retrieval, QA shows reduced heterogeneity between the question and the answer passage/sentence. On one hand, the question is usually in natural language, which is longer than keyword queries and clearer in intent description. On the other hand, the answer passages/sentences are usually much shorter text spans than documents (e.g., the answer passage length of WikiPassageQA data is about 133 words (Cohen, Yang, & Croft, 2018d)), leading to more concentrated topics/semantics. However, vocabulary mismatch is still a basic problem in QA. The notion of relevance is relatively clear in QA, i.e., whether the target passage/sentence answers the question, but assessment is challenging. Ranking models need to capture the patterns expected in the answer passage/sentence based on the intent of the question, such as the matching of the context words, the existence of the expected answer type, and so on.

For the evaluation of QA tasks, several benchmark data sets have been developed, including TREC QA (Voorhees & Tice, 2000), WikiQA (Yang et al., 2015), WebAP (Keikha, Park, & Croft, 2014; Yang et al., 2016b), InsuranceQA (Feng, Xiang, Glass, Wang, & Zhou, 2015), WikiPassageQA (Cohen et al., 2018d) and MS MARCO (Nguyen et al., 2016b). A variety of neural ranking models (Lu &

Li, 2013; Qiu & Huang, 2015; Severyn & Moschitti, 2015; Wang & Nyberg, 2015; Yang, Ai, Guo, & Croft, 2016a) have been tested on these data sets.

2.3. Community question answering

Community question answering (CQA) aims to find answers to users' questions based on existing QA resources in CQA websites, such as Quora [3], Yahoo! Answers [4], Stack Overflow [5], and Zhihu [6]. As a retrieval task, CQA can be further divided into two categories. The first is to directly retrieval answers from the answer pool, which is similar to the above QA task with some additional user behavioral data (e.g., upvotes/downvotes) (Yang et al., 2013). So we will not discuss this format here again. The second is to retrieve similar questions from the question pool, based on the assumption that answers to similar question could answer new questions. Unless otherwise noted, we will refer to the second task format as CQA.

Since it involves the retrieval of similar questions, CQA is significantly different from the previous two tasks due to the homogeneity between the input question and target question. Specifically, both input and target questions are short natural language sentences (e.g. the question length in Yahoo! Answers is between 9 and 10 words on average (Shtok, Dror, Maarek, & Szpektor, 2012)), describing users' information needs. Relevance in CQA refers to semantic equivalence/similarity, which is clear and symmetric in the sense that the two questions are exchangeable in the relevance definition. However, vocabulary mismatch is still a challenging problem as both questions are short and there exist different expressions for the same intent.

For evaluation of the CQA task, a large variety of data sets have been released for research. The well-known data sets include the Quora Dataset[7], Yahoo! Answers Dataset (Qiu & Huang, 2015) and SemEval-2017 Task3 (Nakov et al., 2017). The recent proposed datasets include CQADupStack[8] (Hoogeveen, Verspoor, & Baldwin, 2015), ComQA[9] (Abujabal, Roy, Yahya, & Weikum, 2018) and LinkSO (Liu, Wang, Leng, & Zhai, 2018a). A variety of neural ranking models (Chen et al., 2018b; Pang et al., 2016b; Qiu & Huang, 2015; Wan et al., 2016b; Wang, Hamza, & Florian, 2017) have been tested on these data sets.

2.4. Automatic conversation

Automatic conversation (AC) aims to create an automatic human-computer dialog process for the purpose of question answering, task completion, and social chat (i.e., chit-chat) (Gao, Galley, & Li, 2018). In general, AC could be formulated either as an IR problem that aims to rank/select a proper response from a dialog repository (Ji, Lu, & Li, 2014) or a generation problem that aims to generate an appropriate response with respect to the input utterance (Ritter, Cherry, & Dolan, 2011). In this paper, we restrict AC to the social chat task with the IR formulation, since question answering has already been covered in the above QA task and task completion is usually not taken as an IR problem. From the perspective of conversation context, the IR-based AC could be further divided into single-turn conversation (Wang, Lu, Li, & Chen, 2013) or multi-turn conversation (Wu, Wu, Xing, Zhou, & Li, 2017).

When focusing on social chat, AC also shows homogeneity similar to CQA. That is, both the input utterance and the response are short natural language sentences (e.g., the utterance length of Ubuntu Dialog Corpus is between 10 to 11 words on average and the median conversation length of it is 6 words (Lowe, Pow, Serban, & Pineau, 2015)). Relevance in AC refers to certain semantic correspondence (or coherent structure) which is broad in definition, e.g., given an input utterance "OMG I got myopia at such an 'old' age", the response could range from general (e.g., "Really?") to specific (e.g., "Yeah. Wish a pair of glasses as a gift") (Yan et al., 2016a). Therefore, vocabulary mismatch is no longer the central challenge in AC, as we can see from the example that a good response does not require semantic matching between the words. Instead, it is critical to model correspondence/coherence and avoid general trivial responses.

For the evaluation of different neural ranking models on the AC task, several conversation collections have been collected from social media such as forums, Twitter and Weibo. Specifically, experiments have been conducted over neural ranking models based on collections such as Ubuntu Dialog Corpus (UDC) (Wu et al., 2017; Yang, Zamani, Zhang, Guo, & Croft, 2017; Zhou et al., 2016), Sina Weibo dataset (Wang et al., 2013; Yan et al., 2016a; Yan, Song, Zhou, & Wu, 2016b; Yan, Zhao, & E., 2017), MSDialog (Qu et al., 2018; 2019; Yang et al., 2018) and the "campaign" NTCIR STC (Shang & Sakai).

3. A unified model formulation

Neural ranking models are mostly studied within the LTR framework. In this section, we give a unified formulation of neural ranking models from a generalized view of LTR problems.

Suppose that $\mathcal{S}$ is the *generalized* query set, which could be the set of search queries, natural language questions or input utterances, and $\mathcal{T}$ is the *generalized* document set, which could be the set of documents, answers or responses. Suppose that

[3] https://www.quora.com/
[4] https://answers.yahoo.com
[5] https://www.stackoverflow.com
[6] https://zhihu.com
[7] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs
[8] https://github.com/D1Doris/CQADupStack
[9] http://qa.mpi-inf.mpg.de/comqa

$\mathcal{Y} = \{1, 2, \cdots, l\}$ is the label set where labels represent grades. There exists a total order between the grades $l \succ l-1 \succ \cdots \succ 1$, where $\succ$ denotes the order relation. Let $s_i \in \mathcal{S}$ be the ith query, $T_i = \{t_{i,1}, t_{i,2}, \cdots, t_{i,n_i}\} \in \mathcal{T}$ be the set of documents associated with the query s_i, and $\mathbf{y}_i = \{y_{i,1}, y_{i,2}, \cdots, y_{i,n_i}\}$ be the set of labels associated with query s_i, where n_i denotes the size of T_i and $\mathbf{y}_i$ and $y_{i,j}$ denotes the relevance degree of $t_{i,j}$ with respect to s_i. Let $\mathcal{F}$ be the function class and $f(s_i, t_{i,j}) \in \mathcal{F}$ be a ranking function which associates a relevance score with a query-document pair. Let $L(f; s_i, t_{i,j}, \mathbf{y}_{i,j})$ be the loss function defined on prediction of f over the query-document pair and their corresponding label. So a generalized LTR problem is to find the optimal ranking function f^* by minimizing the loss function over some labeled dataset

$$f^* = \arg\min \sum_i \sum_j L(f; s_i, t_{i,j}, y_{i,j}) \tag{1}$$

Without loss of generality, the ranking function f could be further abstracted by the following unified formulation

$$f(s, t) = g(\psi(s), \phi(t), \eta(s, t)) \tag{2}$$

where s and t are two input texts, ψ, ϕ are representation functions which extract features from s and t respectively, η is the interaction function which extracts features from (s, t) pair, and g is the evaluation function which computes the relevance score based on the feature representations.

Note that for traditional LTR approaches (Liu, 2009), functions ψ, ϕ and η are usually set to be fixed functions (i.e., manually defined feature functions). The evaluation function g can be any machine learning model, such as logistic regression or gradient boosting decision tree, which could be learned from the training data. For neural ranking models, in most cases, all the functions ψ, ϕ, η and g are encoded in the network structures so that all of them can be learned from training data.

In traditional LTR approaches, the inputs s and t are usually raw texts. In neural ranking models, we consider that the inputs could be either raw texts or word embeddings. In other words, embedding mapping is considered as a basic input layer, not included in ψ, ϕ and η.

4. Model architecture

Based on the above unified formulation, here we review existing neural ranking model architectures to better understand their basic assumptions and design principles.

4.1. Symmetric vs. asymmetric architectures

Starting from different underlying assumptions over the input texts s and t, two major architectures emerge in neural ranking models, namely symmetric architecture and asymmetric architecture.

4.1.1. Symmetric architecture

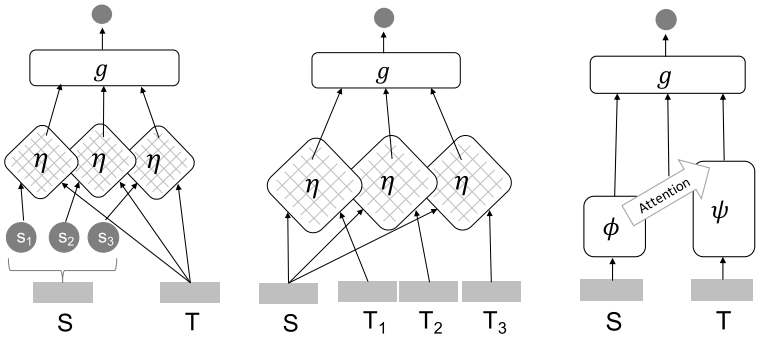
The inputs s and t are assumed to be homogeneous, so that symmetric network structure could be applied over the inputs. Note here symmetric structure means that the inputs s and t can exchange their positions in the input layer without affecting the final output. Specifically, there are two representative symmetric structures, namely siamese networks and symmetric interaction networks.

Siamese networks literally imply symmetric structure in the network architecture. Representative models include DSSM (Huang et al., 2013), CLSM (Shen et al., 2014) and LSTM-RNN (Palangi et al., 2016). For example, DSSM represents two input texts with a unified process including the letter-trigram mapping followed by the multi-layer perceptron (MLP) transformation, i.e., function ϕ is the same as function ψ. After that a cosine similarity function is applied to evaluate the similarity between the two representations, i.e., function g is symmetric. Similarly, CLSM (Shen et al., 2014) replaces the representation functions ψ and ϕ by two identical convolutional neural networks (CNNs) in order to capture the local word order information. LSTM-RNN (Palangi et al., 2016) replaces ψ and ϕ by two identical long short-term memory (LSTM) networks in order to capture the long-term dependence between words.

Symmetric interaction networks, as shown by the name, employ a symmetric interaction function to represent the inputs. Representative models include DeepMatch (Lu & Li, 2013), Arc-II (Hu et al., 2014), MatchPyramid (Pang et al., 2016b) and Match-SRNN (Wan et al., 2016b). For example, Arc-II defines an interaction function η over s and t by computing similarity (i.e., weighted sum) between every n-gram pair from s and t, which is symmetric in nature. After that, several convolutional and max-pooling layers are leveraged to obtain the final relevance score, which is also symmetric over s and t. MatchPyramid defines a symmetric interaction function η between every word pair from s and t to capture fine-grained interaction signals. It then leverages a symmetric evaluation function g, i.e., several 2D CNNs and a dynamic pooling layer, to produce the relevance score.

A similar process can be found in DeepMatch and Match-SRNN.

Symmetric architectures, with the underlying homogeneous assumption, can fit well with the CQA and AC tasks, where s and t usually have similar lengths and similar forms (i.e., both are natural language sentences). They may sometimes work for the ad-hoc retrieval or QA tasks if one only uses document titles/snippets (Huang et al., 2013) or short answer sentences (Yang et al., 2016a) to reduce the heterogeneity between the two inputs.



(a) Query Split (b) Document Split (c) One-way Attention

Fig. 1. Three types of Asymmetric Architecture.

4.1.2. Asymmetric architecture

The inputs s and t are assumed to be heterogeneous, so that asymmetric network structures should be applied over the inputs. Note here asymmetric structure means if we change the position of the inputs s and t in the input layer, we will obtain totally different output. Asymmetric architectures have been introduced mainly in the ad-hoc retrieval task (Huang et al., 2013; Pang et al., 2017), due to the inherent heterogeneity between the query and the document as discussed in Section 2.1. Such structures may also work for the QA task where answer passages are ranked against natural language questions (Dai, Xiong, Callan, & Liu, 2018).

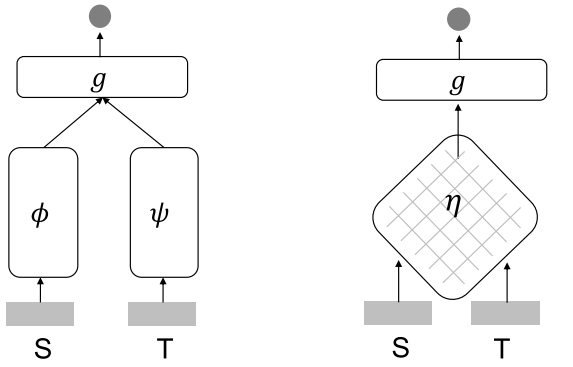
Here we take the ad-hoc retrieval scenario as an example to analyze the asymmetric architecture. We find there are three major strategies used in the asymmetric architecture to handle the heterogeneity between the query and the document, namely query split, document split, and joint split.

- *Query split* is based on the assumption that most queries in ad-hoc retrieval are keyword based, so that we can split the query into terms to match against the document, as illustrated in Fig. 1(a). A typical model based on this strategy is DRMM (Guo et al., 2016). DRMM splits the query into terms and defines the interaction function η as the matching histogram mapping between each query term and the document. The evaluation function g consists of two parts, i.e., a feed-forward network for term-level relevance computation and a gating network for score aggregation. Obviously such a process is asymmetric with respect to the query and the document. K-NRM (Xiong, Dai, Callan, Liu, & Power, 2017b) also belongs to this type of approach. It introduces a kernel pooling function to approximate matching histogram mapping to enable end-to-end learning.
- *Document split* is based on the assumption that a long document could be partially relevant to a query under the scope hypothesis (Robertson & Jones, 1976), so that we split the document to capture fine-grained interaction signals rather than treat it as a whole, as depicted in Fig. 1(b). A representative model based on this strategy is HiNT (Fan et al., 2018). In HiNT, the document is first split into passages using a sliding window. The interaction function η is defined as the cosine similarity and exact matching between the query and each passage. The evaluation function g includes the local matching layers and global decision layers.
- *Joint split*, by its name, uses both assumptions of query split and document split. A typical model based on this strategy is DeepRank (Pang et al., 2017). Specifically, DeepRank splits the document into term-centric contexts with respect to each query term. It then defines the interaction function η between the query and term-centric contexts in several ways. The evaluation function g includes three parts, i.e., term-level computation, term-level aggregation, and global aggregation. Similarly, PACRR (Hui et al., 2017a) takes the query as a set of terms and splits the document using the sliding window as well as the first-k term window.

In addition, in neural ranking models applied for QA, there is another popular strategy leading the asymmetric architecture. We name it *one-way attention mechanism* which typically leverages the question representation to obtain the attention over candidate answer words in order the enhance the answer representation, as illustrated in Fig. 1(c). For example, IARNN (Wang, Liu, & Zhao, 2016) and CompAgg (Wang & Jiang, 2017) get the attentive answer representation sequence that weighted by the question sentence representation.

4.2. Representation-focused vs. interaction-focused architectures

Based on different assumptions over the features (extracted by the representation function ϕ, ψ or the interaction function η) for relevance evaluation, we can divide the existing neural ranking models into another two categories of architectures, namely representation-focused architecture and interaction-focused architecture, as illustrated in Fig. 2. Besides these two basic categories, some neural ranking models adopt a hybrid way to enjoy the merits of both architectures in learning relevance features.



(a) Representation-focused (b) Interaction-focused

Fig. 2. Representation-focused andinteraction-focused architectures.

4.2.1. Representation-focused architecture

The underlying assumption of this type of architecture is that relevance depends on compositional meaning of the input texts. Therefore, models in this category usually define complex representation functions ϕ and ψ (i.e., deep neural networks), but no interaction function η, to obtain high-level representations of the inputs s and t, and uses some simple evaluation function g (e.g. cosine function or MLP) to produce the final relevance score. Different deep network structures have been applied for ϕ and ψ, including fully-connected networks, convolutional networks and recurrent networks.

- To our best knowledge, DSSM (Huang et al., 2013) is the only one that uses the fully-connected network for the functions ϕ and ψ, which has been described in Section 4.1.
- Convolutional networks have been used for ϕ and ψ in Arc-I (Hu et al., 2014), CNTN (Qiu & Huang, 2015) and CLSM (Shen et al., 2014). Take Arc-I as an example, stacked 1D convolutional layers and max pooling layers are applied on the input texts s and t to produce their high-level representations respectively. Arc-I then concatenates the two representations and applies an MLP as the evaluation function g. The main difference between CNTN and Arc-I is the function g, where the neural tensor layer is used instead of the MLP. The description on CLSM could be found in Section 4.1.
- Recurrent networks have been used for ϕ and ψ in LSTM-RNN (Palangi et al., 2016) and MV-LSTM (Wan et al., 2016a). LSTM-RNN uses a one-directional LSTM as ϕ and ψ to encode the input texts, which has been described in Section 4.1. MV-LSTM employs a bi-directional LSTM instead to encode the input texts. Then, the top-k strong matching signals between the two high-level representations are fed to an MLP to generate the relevance score.

By evaluating relevance based on high-level representations of each input text, representation-focused architecture better fits tasks with the global matching requirement (Guo et al., 2016). This architecture is also more suitable for tasks with short input texts (since it is often difficult to obtain good high-level representations of long texts). Tasks with these characteristics include CQA and AC as shown in Section 2. Moreover, models in this category are efficient for online computation, since one can pre-calculate representations of the texts offline once ϕ and ψ have been learned.

4.2.2. Interaction-focused architecture

The underlying assumption of this type of architecture is that relevance is in essence about the relation between the input texts, so it would be more effective to directly learn from interactions rather than from individual representations. Models in this category thus define the interaction function η rather than the representation functions ϕ and ψ, and use some complex evaluation function g (i.e., deep neural networks) to abstract the interaction and produce the relevance score. Different interaction functions have been proposed in literature, which could be divided into two categories, namely non-parametric interaction functions and parametric interaction functions.

- *Non-parametric interaction functions* are functions that reflect the closeness or distance between inputs without learnable parameters. In this category, some are defined over each pair of input word vectors, such as binary indicator function (Pang et al., 2016b; 2017), cosine similarity function (Pang et al., 2016b; 2017; Yang et al., 2016a), dot-product function (Fan et al., 2018; Pang et al., 2016b; 2017) and radial-basis function (Pang et al., 2016b). The others are defined between a word vector and a set of word vectors, e.g. the matching histogram mapping in DRMM (Guo et al., 2016) and the kernel pooling layer in K-NRM (Xiong et al., 2017b).
- *Parametric interaction functions* are adopted to learn the similarity/distance function from data. For example, Arc-II (Hu et al., 2014) uses 1D convolutional layer for the interaction bwteen two phrases. Match-SRNN (Wan et al., 2016b) introduces the neural tensor layer to model complex interactions between input words. Some BERT-based model (Yang, Zhang, & Lin, 2019) takes

attention as the interaction function to learn the interaction vector (i.e., [CLS] vector) between inputs. In general, parametric interaction functions are adopted when there is sufficient training data since they bring the model flexibility at the expense of larger model complexity.

By evaluating relevance directly based on interactions, the interaction-focused architecture can fit most IR tasks in general. Moreover, by using detailed interaction signals rather than high-level representations of individual texts, this architecture could better fit tasks that call for specific matching patterns (e.g., exact word matching) and diverse matching requirement (Guo et al., 2016), e.g., ad-hoc retrieval. This architecture also better fit tasks with heterogeneous inputs, e.g., ad-hoc retrieval and QA, since it circumvents the difficulty of encoding long texts. Unfortunately, models in this category are not efficient for online computation as previous representation-focused models, since the interaction function η cannot be pre-calculated until we see the input pair (s, t). Therefore, a better way for practical usage is to apply these two types of models in a "telescope" setting, where representation-focused models could be applied in an early search stage while interaction-focused models could be applied later on.

It is worth noting that parts of the interaction-focused architectures have some connections to those in the computer vision (CV) area. For example, the designs of MatchPyramid (Pang et al., 2016b) and PACRR (Hui et al., 2017a) are inspired by the neural models for the image recognition task. By viewing the matching matrix as a 2-D image, a CNN network is naturally applied to extract hierarchical matching patterns for relevance estimation. These connections indicate that although neural ranking models are mostly applied over textual data, one may still borrow many useful ideas in neural architecture design from other domains.

4.2.3. Hybrid architecture

In order to take advantage of both representation-focused and interaction-focused architectures, a natural way is to adopt a hybrid architecture for feature learning. We find that there are two major hybrid strategies to integrate the two architectures, namely combined strategy and coupled strategy.

- Combined strategy is a loose hybrid strategy, which simply adopts both representation-focused and interaction-focused architectures as sub-models and combines their outputs for final relevance estimation. A representative model using this strategy is DUET (Mitra et al., 2017). DUET employs a CLSM-like architecture (i.e., a distributed network) and a MatchPyramid-like architecture (i.e., a local network) as two sub-models, and uses a sum operation to combine the scores from the two networks to produce the final relevance score.
- Coupled strategy, on the other hand, is a compact hybrid strategy. A typical way is to learn representations with attention across the two inputs. Therefore, the representation functions ϕ and ψ and the interaction function η are compactly integrated. Representative models using this strategy include IARNN (Wang et al., 2016) and CompAgg (Wang & Jiang, 2017), which have been discussed in the Section 4.1. Both models learn the question and answer representations via some one-way attention mechanism.

4.3. Single-granularity vs. multi-granularity architecture

The final relevance score is produced by the evaluation function g, which takes the features from ϕ, ψ, and η as input for estimation. Based on different assumptions on the estimation process for relevance, we can divide existing neural ranking models into two categories, namely single-granularity models and multi-granularity models.

4.3.1. Single-granularity architecture

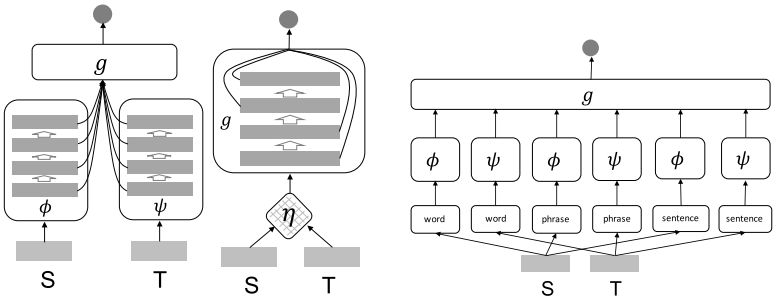
The underlying assumption of the single-granularity architecture is that relevance can be evaluated based on the high-level features extracted by ϕ, ψ and η from the single-form text inputs. Under this assumption, the representation functions ϕ, ψ and the interaction function η are actually viewed as black-boxes to the evaluation function g. Therefore, g only takes their final outputs for relevance computation. Meanwhile, the inputs s and t are simply viewed a set/sequence of words or word embeddings without any additional language structures.

Obviously, the assumption underlying the single-granularity architecture is very simple and basic. Many neural ranking models fall in this category, with either symmetric (e.g., DSSM and MatchPyramid) or asymmetric (e.g., DRMM and HiNT) architectures, either representation-focused (e.g., ARC-I and MV-LSTM) or interaction-focused (e.g., K-NRM and Match-SRNN).

4.3.2. Multi-granularity architecture

The underlying assumption of the multi-granularity architecture is that relevance estimation requires multiple granularities of features, either from different-level feature abstraction or based on different types of language units of the inputs. Under this assumption, the representation functions ϕ, ψ and the interaction function η are no longer black-boxes to g, and we consider the language structures in s and t. We can identify two basic types of multi-granularity, namely vertical multi-granularity and horizontal multi-granularity, as illustrated in Fig. 3.

- *Vertical multi-granularity* takes advantage of the hierarchical nature of deep networks so that the evaluation function g could leverage different-level abstraction of features for relevance estimation. For example, In MultigranCNN (Yin & Schütze, 2015), the representation functions ψ and ϕ are defined as two CNN networks to encode the input texts respectively, and the evaluation function g takes the output of each layer for relevance estimation. MACM (Nie, Sordoni, & Nie, 2018c) builds a CNN over the



(a) Vertical Multi-granularity (b) Horizontal Multi-granularity

Fig. 3. Multi-granularity Architectures.

interaction matrix from η, uses MLP to generate a layer-wise score for each abstraction level of the CNN, and aggregates all the layers' scores for the final relevance estimation. Similar ideas can also be found in MP-HCNN (Rao, Yang, Zhang, Ture, & Lin, 2019) and MultiMatch (Nie, Li, & Nie, 2018a).

- *Horizontal multi-granularity* is based on the assumption that language has intrinsic structures (e.g., phrases or sentences), and we shall consider different types of language units, rather than simple words, as inputs for better relevance estimation. Models in this category typically enhance the inputs by extending it from words to phrases/n-grams or sentences, apply certain single-granularity architectures over each input form, and aggregate all the granularity for final relevance output. For example, in Huang, Yao, Lyu, and Ji (2017), a CNN and an LSTM are applied to obtain the character-level, word-level, and sentence-level representations of the inputs, and each level representations are then interacted and aggregated by the evaluation function g to produce the final relevance score. Similar ideas can be found in Conv-KNRM (Dai et al., 2018) and MIX (Chen et al., 2018a).

As we can see, the multi-granularity architecture is a natural extension of the single-granularity architecture, which takes into account the inherent language structures and network structures for enhanced relevance estimation. With multi-granularity features extracted, models in this category are expected to better fit tasks that require fine-grained matching signals for relevance computation, e.g., ad-hoc retrieval (Dai et al., 2018) and QA (Chen et al., 2018a). However, the enhanced model capability is often reached at the expense of larger model complexity.

5. Model learning

Beyond the architecture, in this section, we review the major learning objectives and training strategies adopted by neural ranking models for comprehensive understadning.

5.1. Learning objective

Similar to other LTR algorithms, the learning objective of neural ranking models can be broadly categorized into three groups: *pointwise, pairwise*, and *listwise*. In this section, we introduce a couple of popular ranking loss functions in each group, and discuss their unique advantages and disadvantages for the applications of neural ranking models in different IR tasks.

5.1.1. Pointwise ranking objective

The idea of pointwise ranking objectives is to simplify a ranking problem to a set of classification or regression problems. Specifically, given a set of query-document pairs $(s_i, t_{i,j})$ and their corresponding relevance annotation $y_{i,j}$, a pointwise learning objective tries to optimize a ranking model by requiring it to directly predict $y_{i,j}$ for $(s_i, t_{i,j})$. In other words, the loss functions of pointwise learning objectives are computed based on each (s, t) pair independently. This can be formulated as

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = \sum_i \sum_j L(y_{i,j}, f(s_i, t_{i,j})) \tag{3}$$

For example, one of the most popular pointwise loss functions used in neural ranking models is *Cross Entropy*:

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = -\sum_i \sum_j y_{i,j} \log(f(s_i, t_{i,j})) + (1 - y_{i,j}) \log(1 - f(s_i, t_{i,j})) \tag{4}$$

where $y_{i,j}$ is a binary label or annotation with probabilistic meanings (e.g., clickthrough rate), and $f(s_i, t_{i,j})$ needs to be rescaled into the range of 0 to 1 (e.g., with a sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$). Example applications include the Convolutional Neural Network for question answering (Severyn & Moschitti, 2015). There are other pointwise loss functions such as *Mean Squared Error* for numerical labels, but they are more commonly used in recommendation tasks.

T he advantages of pointwise ranking objectives are two-fold. First, pointwise ranking objectives are computed based on each

query-document pair (s_i, $t_{i,j}$) separately, which makes it simple and easy to scale. Second, the outputs of neural models learned with pointwise loss functions often have real meanings and value in practice. For instance, in sponsored search, a model learned with cross entropy loss and clickthrough rates can directly predict the probability of user clicks on search ads, which is more important than creating a good result list in some application scenarios.

In general, however, pointwise ranking objectives are considered to be less effective in ranking tasks. Because pointwise loss functions consider no document preference or order information, they do not guarantee to produce the best ranking list when the model loss reaches the global minimum. Therefore, better ranking paradigms that directly optimize document ranking based on pairwise loss functions and listwise loss functions have been proposed for LTR problems.

5.1.2. Pairwise ranking objective

Pairwise ranking objectives focus on optimizing the relative preferences between documents rather than their labels. In contrast to pointwise methods where the final ranking loss is the sum of loss on each document, pairwise loss functions are computed based on the permutations of all possible document pairs (Chen, Liu, Lan, Ma, & Li, 2009). It usually can be formalized as

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = \sum_{i} \sum_{(j,k), y_{i,j} \succ y_{i,k}} L(f(s_i, t_{i,j}) - f(s_i, t_{i,k})) \tag{5}$$

where $t_{i,j}$ and $t_{i,k}$ are two documents for query s_i and $t_{i,j}$ is preferable comparing to $t_{i,k}$ (i.e., $y_{i,j} \succ y_{i,k}$). For instance, a well-known pairwise loss function is *Hingle loss*:

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = \sum_{i} \sum_{(j,k), y_{i,j} \succ y_{i,k}} \max(0, 1 - f(s_i, t_{i,j}) + f(s_i, t_{i,k})) \tag{6}$$

Hingle loss has been widely used in the training of neural ranking models such as DRMM (Guo et al., 2016) and K-NRM (Xiong et al., 2017b). Another popular pairwise loss function is the pairwise cross entropy defined as

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = -\sum_{i} \sum_{(j,k), y_{i,j} \succ y_{i,k}} \log \sigma(f(s_i, t_{i,j}) - f(s_i, t_{i,k})) \tag{7}$$

where $\sigma(x) = \frac{1}{1 + \exp(-x)}$. Pairwise cross entropy is first proposed in RankNet by Burges et al. (2005), which is considered to be one of the initial studies on applying neural network techniques to ranking problems.

Ideally, when pairwise ranking loss is minimized, all preference relationships between documents should be satisfied and the model will produce the optimal result list for each query. This makes pairwise ranking objectives effective in many tasks where performance is evaluated based on the ranking of relevant documents. In practice, however, optimizing document preferences in pairwise methods does not always lead to the improvement of final ranking metrics due to two reasons: (1) it is impossible to develop a ranking model that can correctly predict document preferences in all cases; and (2) in the computation of most existing ranking metrics, not all document pairs are equally important. This means that the performance of pairwise preference prediction is not equal to the performance of the final retrieval results as a list. Given this problem, previous studies (Ai, Bi, Guo, & Croft, 2018a; Burges, 2010; Taylor, Guiver, Robertson, & Minka, 2008; Xia, Liu, Wang, Zhang, & Li, 2008) further proposed listwise ranking objectives for learning to rank.

5.1.3. Listwise ranking objective

The idea of listwise ranking objectives is to construct loss functions that directly reflect the model's final performance in ranking. Instead of comparing two documents each time, listwise loss functions compute ranking loss with each query and their candidate document list together. Formally, most existing listwise loss functions can be formulated as

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = \sum_{i} L(\{y_{i,j}, f(s_i, t_{i,j}) | t_{i,j} \in \mathcal{T}_i\}) \tag{8}$$

where $\mathcal{T}_i$ is the set of candidate documents for query s_i. Usually, L is defined as a function over the list of documents sorted by $y_{i,j}$, which we refer to as π_i, and the list of documents sorted by $f(s_i, t_{i,j})$. For example, Xia et al. (2008) proposed *ListMLE* for listwise ranking as

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = \sum_{i} \sum_{j=1}^{|\pi_i|} \log P(y_{i,j} | \mathcal{T}_i^{(j)}, f) \tag{9}$$

where $P(y_{i,j} | \mathcal{T}_i^{(j)}, f)$ is the probability of selecting the jth document in the optimal ranked list π_i with f:

$$P(y_{i,j} | \mathcal{T}_i^{(j)}, f) = \frac{\exp(f(s_i, t_{i,j}))}{\sum_{k=j}^{|\pi_i|} \exp(f(s_i, t_{i,k}))} \tag{10}$$

Intuitively, ListMLE is the log likelihood of the optimal ranked list given the current ranking function f, but computing log likelihood on all the result positions is computationally prohibitive in practice. Thus, many alternative functions have been proposed for listwise ranking objectives in the past ten years. One example is the *Attention Rank* function used in the Deep Listwise Context Model proposed by Ai et al. (2018a):

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = -\sum_i \sum_j P(t_{i,j}|\mathcal{Y}_i, \mathcal{T}_i) \log P(t_{i,j}|f, \mathcal{T}_i)$$
$$\text{where} \quad P(t_{i,j}|\mathcal{Y}_i, \mathcal{T}_i) = \frac{\exp(y_{i,j})}{\sum_{k=1}^{|\mathcal{T}_i|} \exp(y_{i,k})},$$
$$P(t_{i,j}|fi, \mathcal{T}_i) = \frac{\exp(f(s_i, t_{i,j}))}{\sum_{k=1}^{|\mathcal{T}_i|} \exp(f(s_i, t_{i,k}))} \tag{11}$$

When the labels of documents (i.e., $y_{i,j}$) are binary, we can further simplify the Attention Rank function with a softmax cross entropy function as

$$L(f; \mathcal{S}, \mathcal{T}, \mathcal{Y}) = -\sum_i \sum_j y_{i,j} \log \frac{\exp(f(s_i, t_{i,j}))}{\sum_{k=1}^{|\mathcal{T}_i|} \exp(f(s_i, t_{i,k}))} \tag{12}$$

The softmax-based listwise ranking loss is one of the most popular learning objectives for neural ranking models such as GSF (Ai, Wang, Golbandi, Bendersky, & Najork, 2018c). It is particularly useful when we train neural ranking models with user behavior data (e.g., clicks) under the unbiased learning framework (Ai, Mao, Liu, & Croft, 2018b). There are other types of listwise loss functions proposed under different ranking frameworks in the literature (Burges, 2010; Taylor et al., 2008). We ignore them in this paper since they are not popular in the studies of neural IR.

While listwise ranking objectives are generally more effective than pairwise ranking objectives, their high computational cost often limits their applications. They are suitable for the re-ranking phase over a small set of candidate documents. Since many practical search systems now use neural models for document re-ranking, listwise ranking objectives have become increasingly popular in neural ranking frameworks (Ai et al., 2018a; Ai et al., 2018b; Ai et al., 2018c; Huang et al., 2013; Mitra et al., 2017; Shen et al., 2014).

5.1.4. Multi-task learning objective

In some cases, the optimization of neural ranking models may include the learning of multiple ranking or non-ranking objectives at the same time. The motivation behind this approach is to use the information from one domain to help the understanding of information from other domains. For example, Liu et al. (2015) proposed to unify the representation learning process for query classification and Web search by training a deep neural network in which the final layer of hidden variables are used to optimize both a classification loss and a ranking loss. Chapelle et al. (2010) proposed a multi-boost algorithm to simultaneously learn ranking functions based on search data collected from 15 countries.

In general, the most common methodology used by existing multi-task learning algorithms is to construct shared representations that are universally effective for ranking in multiple tasks or domains. To do so, previous studies mostly focus on constructing regularizations or restrictions on model optimizations so that the final model is not specifically designed for a single ranking objective (Chapelle et al., 2010; Liu et al., 2015). Inspired by recent advances on generative adversarial networks (GAN) (Cohen, Mitra, Hofmann, & Croft, 2018b; Goodfellow et al., 2014) introduced an adversarial learning framework that jointly learns a ranking function with a discriminator which can distinguish data from different domains. By training the ranking function to produce representations that cannot be discriminated by the discriminator, they teach the ranking system to capture domain-independent patterns that are usable in cross-domain applications. This is important as it can significantly alleviate the problem of data sparsity in specific tasks and domains.

5.2. Training strategies

Given the data available for training a neural ranking model, an appropriate training strategy should be chosen. In this section, we briefly review a set of effective training strategies for neural ranking models, including supervised, semi-supervised, and weakly supervised learning.

Supervised learning refers to the most common learning strategy in which query-document pairs are labeled. The data can be labeled by expert assessors, crowdsourcing, or can be collected from the user interactions with a search engine as implicit feedback. In this training strategy, it is assumed that a sufficient amount of labeled training data is available. Given this training strategy, one can train the model using any of the aforementioned learning objectives, e.g., pointwise and pairwise. However, since neural ranking models are usually data "hungry", academic researchers can only learn models with constrained parameter spaces under this training paradigm due to the limited annotated data. This has motivated researchers to study learning from limited data for information retrieval (Zamani, Dehghani, Diaz, Li, & Craswell, 2018c).

Weakly supervised learning refers to a learning strategy in which the query-document labels are automatically generated using an existing retrieval model, such as BM25. The use of pseudo-labels for training ranking models has been proposed by Asadi, Metzler, Elsayed, and Lin (2011). More recently, Dehghani et al. (2017b) proposed to train neural ranking models using weak supervision and observed up to 35% improvement compared to BM25 which plays the role of weak labeler. This learning strategy does not require labeled training data. In addition to ranking, weak supervision has shown successful results in other information retrieval tasks, including query performance prediction (Zamani, Croft, & Culpepper, 2018a), learning relevance-based word embedding (Zamani & Croft, 2017), and efficient learning to rank (Cohen, Foley, Zamani, Allan, & Croft, 2018a).

Semi-supervised learning refers to a learning strategy that leverages a small set of labeled query-document pairs plus a large set of unlabeled data. Semi-supervised learning has been extensively studied in the context of learning to rank. Preference

regularization (Szummer & Yilmaz, 2011), feature extraction using KernelPCA (Duh & Kirchhoff, 2008), and pseudo-label generation using labeled data (Zhang, He, & Luo, 2016) are examples of such approaches. In the realm of neural models, fine-tuning weak supervision models using a small set of labeled data (Dehghani et al., 2017b) and controlling the learning rate in learning from weakly supervised data using a small set of labeled data (Dehghani, Severyn, Rothe, & Kamps, 2017a) are another example of semi-supervised approaches to ranking. Recently, Li, Cheng, and Jia (2018a) proposed a neural model with a joint supervised and unsupervised loss functions. The supervised loss accounts for the error in query-document matching, while the unsupervised loss computes the document reconstruction error (i.e., auto-encoders).

6. Model comparison

In this section, we compare the empirical evaluation results of the previously reviewed neural ranking models on several popular benchmark data sets. We mainly survey and analyze the published results of neural ranking models for the ad-hoc retrieval and QA tasks. Note that sometimes it is difficult to compare published results across different papers–small changes such as different tokenization, stemming, etc. can lead to significant differences. Therefore, we attempt to collect results from papers that contain comparisons across some of these models performed at a single site for fairness.

6.1. Empirical comparison on ad-hoc retrieval

To better understand the performances of different neural ranking models on ad-hoc retrieval, we show the published experimental results on benchmark datasets. Here, we choose three representative datasets for ad-hoc retrieval: (1) Robust04 dataset is a standard ad-hoc retrieval dataset where the queries are from TREC Robust Track 2004. (2) Gov2_{MQ2007} is an Web Track ad-hoc retrieval dataset where the collection is the Gov2 corpus. The queries are from the Million Query Track of TREC 2007. (3) Sougou-Log dataset (Xiong et al., 2017b) is built on query logs sampled from search logs of Sougou.com. (4) WT09-14 is the 2009–2014 TREC Web Track, which are based on the ClueWeb09 and ClueWeb12 datasets. The detailed data statistics can be found in related literature (Fan et al., 2018; Guo et al., 2016; Hui, Yates, Berberich, & de Melo, 2018; Pang et al., 2017; Xiong et al., 2017b).

For meaningful comparison, we have tried our best to restrict the reported results to be under the same experimental settings. Specifically, experiments on Robust04 take the title as the query, and all the documents are processed with the Galago Search Engine[10] (Guo et al., 2016; Zamani et al., 2018b). For experiments on the Gov2_{MQ2007} dataset, all the queries and documents are processed using the Galago Search Engine under the same setting as described in (Fan et al., 2018; Pang et al., 2017). Besides, the results on the WT09-14 dataset and the Sougou-Log dataset are all from a same paper (Dai et al., 2018; Hui et al., 2018) respectively.

Table 1 shows an overview of previous published results on ad-hoc retrieval datasets. We have included some well-known probabilistic retrieval models, pseudo-relevance feedback (PRF) models and LTR models as baselines. Based on the results, we have the following observations:

1. The probabilistic models (i.e., QL and BM25), although simple, can already achieve reasonably good performance. The traditional PRF model (i.e., RM3) and LTR models (i.e., RankSVM and LambdaMart) with human designed features are strong baselines whose performance is hard to beat for most neural ranking models based on raw texts. However, the PRF technique can also be leveraged to enhance neural ranking models (e.g., SNRM + PRF (Zamani et al., 2018b) and NPRF + DRMM (Li et al., 2018b) in Table 1), while human designed LTR features can be integrated into neural ranking models (Fan et al., 2017a; Pang et al., 2017) to improve the ranking performance.
2. There seems to be a paradigm shift of the neural ranking model architectures from symmetric to asymmetric and from representation-focused to interaction-focused over time. This is consistent with our previous analysis where asymmetric and interaction-focused structures may fit better with the ad-hoc retrieval task which shows heterogeneity inherently.
3. With bigger data size in terms of distinct number of queries and labels (i.e., Sogou-Log $\succ$ GOV2_{MQ2007} $\succ$ WT09-14 $\succ$ Robust04), neural models are more likely to achieve larger performance improvement against non-neural models. As we can see, the best neural models based on raw texts can significantly outperform LTR models with human designed features on Sogou-Log dataset.
4. Based on the reported results, in general, we observe that the asymmetric, interaction-focused, multi-granularity architecture can work better than the symmetric, representation-focused, single-granularity architecture on the ad-hoc retrieval tasks. There is one exception, i.e., SNRM on Robust04. However, this model was trained with a large amount of data using the weak supervision strategy, and may not be appropriate to directly compare with those models trained on Robust04 alone.

6.2. Empirical comparison on QA

In order to understand the performance of different neural ranking models reviewed in this paper for the QA task, we survey the previously published results on three QA data sets, including TREC QA (Wang, Smith, & Mitamura, 2007), WikiQA (Yang et al., 2015) and Yahoo! Answers (Wan et al., 2016a). TREC QA and WikiQA are answer sentence selection/retrieval data sets and they mainly contain factoid questions, while Yahoo! Answers is an answer passage retrieval data set sampled from the CQA website Yahoo! Answers. The detailed data statistics can be found in related literature (Wan et al., 2016a; Yang et al., 2015; Yu, Hermann, Blunsom,

[10] http://www.lemurproject.org/galago.php

Table 1
Overview of previously published results on ad hoc retrieval datasets. The citation in each row denotes the original paper where the method is proposed. The superscripts 1–6 denote that the results are cited from Dai et al. (2018); Fan et al. (2018); Guo et al. (2016); Hui et al. (2018); Li et al. (2018b); Pang et al. (2017); Zamani et al. (2018b), respectively. The subscripts denote the model architecture belongs to (S)ymmetric or (A)symmetric/(R)epresentation-focused or (I)nteraction-focused or (H)ybrid/Singe-(G)ranularity or (M)ulti-granularity. The back slash symbols denote that there are no published results for the specific model on the specific data set in the related literature.

Model \ DataSet	Robust04		GOV2$_{MQ2007}$		WT09-14	Sougo-Log
	MAP	P@20	MAP	P@10	ERR@20	NDCG@1
BM25 (Robertson & Walker, 1994) (1994)^{1,2}	0.255	0.370	0.450	0.366	\	0.142
QL (Ponte & Croft, 1998) (1998)^{1,4}	0.253	0.369	\	\	0.113	0.126
RM3 (Lavrenko & Croft, 2001)(2001)^{5}	0.287	0.377	\	\	\	\
RankSVM (Joachims, 2002) (2002)^{2}	\	\	0.464	0.381	\	0.146
LambdaMart (Burges, 2010) (2010)^{2}	\	\	0.468	0.384	\	\
DSSM (Huang et al., 2013) (2013)$^{1,2}_{S/R/G}$	0.095	0.171	0.409	0.352	\	\
CDSSM (Shen et al., 2014) (2014)$^{1,2}_{S/R/G}$	0.067	0.125	0.364	0.291	\	0.144
ARC-I (Hu et al., 2014) (2014)$^{1,2}_{S/R/G}$	0.041	0.065	0.417	0.364	\	\
ARC-II (Hu et al., 2014) (2014)$^{1,2}_{S/I/G}$	0.067	0.128	0.421	0.366	\	\
MP (Pang et al., 2016b) (2016)$^{1,2,4}_{S/I/G}$	0.189	0.290	0.434	0.371	0.148	0.218
Match-SRNN (Wan et al., 2016b) (2016)$^{2}_{S/H/G}$	\	\	0.456	0.384	\	\
DRMM (Guo et al., 2016) (2016)$^{1,2,4}_{A/I/G}$	0.279	0.382	0.467	0.388	0.171	0.137
Duet (Mitra et al., 2017) (2017)$^{3,4}_{A/H/G}$	\	\	0.474	0.398	0.134	\
DeepRank (Pang et al., 2017) (2017)$^{2}_{A/I/G}$	\	\	0.497	0.412	\	\
K-NRM (Xiong et al., 2017b) (2017)$^{4}_{A/I/G}$	\	\	\	\	0.154	0.264
PACRR (Hui, Yates, Berberich, & de Melo, 2017b) (2017)$^{6,4}_{A/I/M}$	0.254	0.363	\	\	0.191	\
Co-PACRR (Hui et al., 2018) (2018)$^{4}_{A/I/M}$	\	\	\	\	0.201	\
SNRM (Zamani et al., 2018b) (2018)$^{5}_{S/R/G}$	0.286	0.377	\	\	\	\
SNRM + PRF (Zamani et al., 2018b) (2018)$^{5}_{S/R/G}$	0.297	0.395	\	\	\	\
CONV-KNRM (Dai et al., 2018) (2018)$^{4}_{A/I/M}$	\	\	\	\	\	0.336
NPRF-KNRM (Li et al., 2018b) (2018)$^{6}_{A/I/G}$	0.285	0.393	\	\	\	\
NPRF-DRMM (Li et al., 2018b) (2018)$^{6}_{A/I/G}$	0.290	0.406	\	\	\	\
HiNT (Fan et al., 2018) (2018)$^{3}_{A/I/G}$	\	\	0.502	0.418		\

& Pulman, 2014).

We have tried our best to report results under the same experimental settings for fair comparison between different methods. Specifically, the results on TREC QA are over the raw version of the data (Rao, He, & Lin, 2016)[11]

Table 2 shows the overview of the published results on the QA benchmark data sets. We include several traditional non-neural methods as baselines. We summarize our observations as follows:

1. Unlike ad-hoc retrieval, symmetric architectures have been more widely adopted in the QA tasks possibly due to the increased homogeneity between the question and the answer, especially for answer sentence retrieval data sets like TREC QA and WikiQA.
2. Representation-focused architectures have been more adopted on short answer sentence retrieval data sets, i.e., TREC QA and WikiQA, while interaction-focused architectures have been more adopted on longer answer passage retrieval data sets, e.g., Yahoo! Answer. However, unlike ad-hoc retrieval, there seems to be no clear winner between the representation-focused architecture and the interaction-focused architecture on QA tasks.
3. Similar to ad-hoc retrieval, neural models are more likely to achieve larger performance improvement against non-neural models on bigger data sets. For example, on small data set like TREC QA, feature engineering based methods such as LCLR can achieve very strong performance. However, on large data set like WikiQA and Yahoo! Answers, we can see a clear gap between neural models and non-neural models.
4. The performance in general increases over time, which might be due to the increased model capacity as well as the adoption of some advanced approaches, e.g., the attention mechanism. For example, IARNN utilizes attention-based RNN models with GRU to get an attentive sentence representation. MIX extracts grammar information and integrates attention matrices in the attention channels to encapsulate rich structural patterns. aNMM adopts attention mechanism to encode question term importance for aggregating interaction matching features.

[11] https://aclweb.org/aclwiki/Question_Answering_(State_of_the_art). WikiQA only has a single version with the same train/ valid/ test data partitions (Yang et al., 2015). Yahoo Answers data is the processed version from the same related work (Wan et al., 2016a). Therefore, questions and answer candidates in all the train/valid/test sets used in different surveyed papers are the same, and the results are comparable with each other.

Table 2
Overview of previously published results on QA benchmark data sets. The citation in each row denotes the original paper where the method is proposed. The superscripts 1–10 denote that the results are cited from Chen et al. (2018a); Tay, Phan, Luu, and Hui (2017); Tay, Tuan, and Hui (2018); Wan et al. (2016a,b); Wang et al. (2016); Wang and Jiang (2017); Yang et al. (2016a, 2015); Yu et al. (2014), respectively. The subscripts denote the model architecture belongs to (S)ymmetric or (A)symmetric/(R)epresentation-focused or (I)nteraction-focused or (H)ybrid/Single-(G)ranularity or (M)ulti-granularity. The back slash symbols denote that there are no published results for the specific model on the specific data set in the related literature.

Data Set	TREC QA		WikiQA		Yahoo! Answers	
Model	MAP	MRR	MAP	MRR	P@1	MRR
BM25 (Robertson & Walker, 1994) $(1994)^{2}$	\	\	\	\	0.579	0.726
LCLR (Yih, Chang, Meek, & Pastusiak, 2013) $(2013)^{1,9}$	0.709	0.770	0.599	0.609		\
Word Cnt (Yu et al., 2014) $(2014)^{1,9}$	0.571	0.627	0.489	0.492		\
Wgt Word Cnt (Yu et al., 2014) $(2014)^{1,9}$	0.596	0.652	0.510	0.513		\
DeepMatch (Lu & Li, 2013) $(2013)^{5}_{S/I/G}$	\	\	\	\	0.452	0.679
CNN (Yu et al., 2014) $(2014)^{1,9}_{S/R/G}$	0.569	0.661	0.619	0.628	\	\
CNN-Cnt (Yu et al., 2014) $(2014)^{1,9}_{S/R/G}$	0.711	0.785	0.652	0.665	\	\
ARC-I (Hu et al., 2014) $(2014)^{2}_{S/R/G}$	\	\	\	\	0.581	0.756
ARC-II (Hu et al., 2014) $(2014)^{2}_{S/I/G}$	\	\	\	\	0.591	0.765
CDNN (Severyn & Moschitti, 2015) $(2015)^{3}_{S/R/G}$	0.746	0.808	\	\	\	\
BLSTM (Wang & Nyberg, 2015) $(2015)^{3}_{S/R/G}$	0.713	0.791	\	\	\	\
CNTN (Qiu & Huang, 2015) $(2015)^{2,6}_{S/R/G}$	0.728	0.783	\	\	0.626	0.781
MultiGranCNN (Yin & Schütze, 2015) $(2015)^{2}_{S/I/M}$	\	\	\	\	0.725	0.840
LSTM-RNN (Palangi et al., 2016) $(2016)^{2}_{S/R/G}$	\	\	\	\	0.690	0.822
MV-LSTM (Wan et al., 2016a) $(2016)^{2,6}_{S/R/G}$	0.708	0.782	\	\	0.766	0.869
MatchPyramid (Pang et al., 2016b) $(2016)^{2}_{S/I/G}$	\	\	\	\	0.764	0.867
aNMM (Yang et al., 2016a) $(2016)^{3}_{A/I/G}$	0.750	0.811	\	\	\	\
Match-SRNN (Wan et al., 2016b) $(2016)^{2}_{S/I/G}$		\	\	\	0.790	0.882
IARNN (Wang et al., 2016) $(2016)^{4}_{A/H/G}$	\	\	0.734	0.742	\	\
HD-LSTM (Tay et al., 2017) $(2017)^{6}_{S/R/G}$	0.750	0.815	\	\	\	\
CompAgg (Wang & Jiang, 2017) $(2017)^{7}_{A/I/G}$	\	\	0.743	0.755	\	\
HyperQA (Tay et al., 2018) $(2018)^{8}_{S/R/G}$	0.770	0.825	0.712	0.727	\	\
MIX (Chen et al., 2018a) $(2018)^{10}_{S/I/M}$	\	\	0.713	\	\	\

7. Trending topics

In this section, we discuss several trending topics related to neural ranking models. Some of these topics are important but have not been well addressed in this field, while some are very promising directions for future research.

7.1. Indexing: From re-ranking to ranking

Modern search engines take advantage of a multi-stage cascaded architecture in order to efficiently provide accurate result lists to users. In more detail, there can be a stack of rankers, starting from an efficient high-recall model. Learning to rank models are often employed to model the last stage ranker whose goal is to re-rank a small set of documents retrieved by the early stage rankers. The main objective of these learning to rank models is to provide high-precision results.

Such a multi-stage cascaded architecture suffers from an error propagation problem. In other words, the errors initiated by the early stage rankers are propagated to the last stage. This clearly shows that multi-stage systems are not optimal. However, for efficiency reasons, learning to rank models cannot be used as the sole ranker to retrieve from large collections, which is a disadvantage for such models.

To address this issue, Zamani et al. (2018b) recently argued that the sparse nature of natural languages enables efficient term-matching retrieval models to take advantage of an *inverted index* data structure for efficient retrieval. Therefore, they proposed a standalone neural ranking model (SNRM) that learns high-dimensional sparse representations for queries and documents. In more detail, this type of model should optimize two objectives: (*i*) a relevance objective that maximizes the effectiveness of the model in terms of the retrieval performance, and (*ii*) a sparsity objective that is equivalent to minimizing L_0 of the query and document representations. SNRM has shown superior performance compared to competitive baselines and has performed as efficiently as term-matching models, such as TF-IDF and BM25.

Learning inverted indexes has been also started to be explored in the database community. Kraska, Beutel, Chi, Dean, and Polyzotis (2018) recently proposed to look at indexes as models. For example, a B-Tree-Index can be seen as a function that maps each key to a position of record in a sorted list. They proposed to replace traditional indexes used in databases with the indexes learned

using deep learning technologies. Their models demonstrate a significant conflict reduction and memory footprint improvement.

Graph-based hashing and indexing algorithms have also attracted a considerable attention, which could be leveraged to index neural representations for the initial retrieval. For instance, Boytsov, Novak, Malkov, and Nyberg (2016) proposed to replace term-matching retrieval models with approximate nearest neighbor algorithms. Gysel, de Rijke, and Kanoulas (2018) used a similar idea to design an unsupervised neural retrieval model, however, their model architecture is not scalable to large document collections.

Moving from re-ranking a small set of documents to retrieving documents from a large collection is a recent research direction with a number of unanswered questions that require further investigation. For example, understanding and interpreting the learned neural representations has yet to be addressed. Furthermore, there is a known trade-off between efficiency and effectiveness in information retrieval systems, however, understanding this trade-off in learning inverted indexes requires further research. In addition, although index compression is a common technique in the search engine industry to reduce the size of the posting lists and improve efficiency, compression of the learned latent indexes is an unexplored area of research.

In summary, learning to index and developing effective and at the same time efficient retrieval models is a promising direction in neural IR research, however, we still face several open questions in this area.

7.2. Learning with external knowledge

Most existing neural ranking models focus on learning the matching patterns between the two input texts. In recent years, some researchers have gone beyond matching textual objects by leveraging external knowledge to enhance the ranking performance. These research works can be grouped into two categories: (1) learning with external structured knowledge such as knowledge bases (Liu, Xiong, Sun, & Liu, 2018b; Nguyen, Tamine, Soulier, & Bricon-Souf, 2016a; Shen et al., 2018; Song et al., 2018; Xiong et al., 2017a; Xu, Liu, Wang, Sun, & Wang, 2016); (2) learning with external unstructured knowledge such as retrieved top results, topics or tags (Ghazvininejad et al., 2018; Wu, Wu, Xu, & Li, 2018; Yang et al., 2018). We now briefly review this work.

The first category of research explored improving neural ranking models with semantic information from knowledge bases. Liu et al. (2018b) proposed EDRM that incorporates entities in interaction-focused neural ranking models. EDRM first learns the distributed representations of entities using their semantics from knowledge bases in descriptions and types. Then the model matches documents to queries with both bag-of-words and bag-of-entities. Similar approaches were proposed by Xiong et al. (2017a), which also models queries and documents with word-based representations and entity-based representations. Nguyen et al. (2016a) proposed combining distributional semantics learned through neural networks and symbolic semantics held by extracted concepts or entities from text knowledge bases to enhance the learning algorithm of latent representations of queries and documents. Shen et al. (2018) proposed the KABLSTM model, which leverages external knowledge from knowledge graphs to enrich the representational learning of QA sentences. Xu et al. (2016) designed a Recall gate, where domain knowledge can be transformed into the extra global memory of LSTM, with the aim of enhancing LSTM by cooperating with its local memory to capture the implicit semantic relevance between sentences within conversations.

Beyond structured knowledge in knowledge bases, other research has explored how to integrate external knowledge from unstructured texts, which are more common for information on the Web. Yang et al. (2018) studied response ranking in information-seeking conversations and proposed two effective methods to incorporate external knowledge into neural ranking models with pseudo-relevance feedback (PRF) and QA correspondence knowledge distillation. They proposed to extract the “correspondence” regularities between question and answer terms from retrieved external QA pairs as external knowledge to help response selection. Another representative work on integrating unstructured knowledge into neural ranking models is the KEHNN model proposed by Wu et al. (2018), which defined prior knowledge as topics, tags, and entities related to the text pair. KEHNN represents global context obtained from external textual collection, and then exploits a knowledge gate to fuse the semantic information carried by the prior knowledge into the representation of words. Finally, it generates a knowledge enhanced representation for each word to construct the interaction matrix between text pairs.

In summary, learning with external knowledge is an active research area related to neural ranking models. More research efforts are needed to improve the effectiveness of neural ranking models with distilled external knowledge and to understand the role of external knowledge in ranking tasks.

7.3. Learning with visualized technology

We have discussed many neural ranking models in this survey under the textual IR scenario. There have also been a few studies showing that the textual IR problem could be solved visually. The key idea is that we can construct the matching between two inputs as an image so that we can leverage deep neural models to estimate the relevance based on visual features. The advantage of the matching image, compared with traditional matching matrix, is that it can keep the layout information of the original inputs so that many useful features such as spatial proximity, font size and colors could be modeled for relevance estimation. This is especially useful when we consider ad-hoc retrieval tasks on the Web where pages are often well designed documents with rich layout information.

Specifically, Fan et al. (2017a) proposed a visual perception model (ViP) to perceive visual features for relevance estimation. They first rendered the Web pages into query-independent snapshots and query-dependent snapshots. Then, the visual features are learned through a combination of CNN and LSTM, inspired by users’ reading behaviour. The results have demonstrated the effectiveness of learning the visual features of document for ranking problems. Zhang, Liu, Ma, and Tian (2018) proposed a joint relevance estimation model which learns visual patterns, textual semantics and presentation structures jointly from screenshots, titles, snippets and HTML

source codes of search results. Their results have demonstrated the viability of the visual features in search result page relevance estimation. Recently, Akker, Markov, and de Rijke (2019) built a dataset for the LTR task with visual features, named Visual learning TO Rank (ViTOR). The ViTOR dataset consists of visual snapshots, non-visual features and relevance judgments for ClueWeb12 webpages and TREC Web Track queries. Their results have demonstrated that visual features can significantly improve the LTR performance.

In summary, solving the textual ranking problem through visualized technology is a novel and interesting direction. In some sense, this approach simulates human behavior as we also judge relevance through visual perception. The existing work has only demonstrated the effectiveness of visual features in some relevance assessment tasks. However, more research is needed to understand what can be learned by such visualized technology beyond those text-based methods, and what IR applications could benefit from such models.

7.4. Learning with context

Search queries are often short and cannot precisely express the underlying information needs. To address this issue, a common strategy is to exploit *query context* to improve the retrieval performance. Different types of query context have been explored in the literature:

- Short-term history: the user past interactions with the system in the current search session (Shen, Tan, & Zhai, 2005; Ustinovskiy & Serdyukov, 2013; Xiang et al., 2010).
- Long-term history: the historical information of the user's queries that is often used for web search personalization (Bennett et al., 2012; Matthijs & Radlinski, 2011).
- Situational context: the properties of the current search request, independent from the query content, such as location and time (Bennett, Radlinski, White, & Yilmaz, 2011; Zamani, Bendersky, Wang, & Zhang, 2017).
- (Pseudo-) relevance feedback: explicit, implicit, or pseudo relevance signals for a given query can be used as the query context to improve the retrieval performance.

Although query context has been widely explored in the literature, incorporating query context into neural ranking models is relatively less studied. Zamani et al. (2017) proposed a deep and wide network architecture in which the deep part of the model learns abstract representations for contextual features, while the wide part of the model uses raw contextual features in binary format in order to avoid information loss as a result of high-level abstraction. Ahmad, Chang, and Wang (2018) incorporated short-term history information into a neural ranking model by multi-task training of document ranking and query suggestion. Short- and long-term history have been also used by Chen, Cai, Chen, and de Rijke (2018c) for query suggestion.

In addition, learning high-dimensional representation for pseudo-relevance feedback has been also studied in the literature. In this area, embedding-based relevance models (Zamani & Croft, 2016) extend the original relevance models (Lavrenko & Croft, 2001) by considering word embedding vectors. The word embedding vectors can be obtained from self-supervised algorithms, such as word2vec (Mikolov, Chen, Corrado, & Dean, 2013a), or weakly supervised algorithms, such as relevance-based word embedding (Zamani & Croft, 2017). Zamani, Dadashkarimi, Shakery, and Croft (2016) proposed RFMF, the first pseudo-relevance feedback model that learns latent factors from the top retrieved document. RFMF uses non-negative matrix factorization for learning latent representations for words, queries, and documents. Later on, Li et al. (2018b) extended existing neural ranking models, e.g., DRMM (Guo et al., 2016) and KNRM (Xiong et al., 2017b), by a neural pseudo-relevance feedback approach, called NPRF. The authors showed that in many cases extending a neural ranking model with NPRF leads to significant improvements. Zamani et al. (2018b) also made a similar conclusion by extending SNRM with pseudo-relevance feedback.

In summary, with the emergence of interactive or conversational search system, context-aware ranking would be an indispensable technology in these scenarios. These exist several open research questions on how to incorporate query context information in neural ranking models. More research work is expected in this direction in the short future.

7.5. Neural ranking model understanding

Deep learning techniques have been widely criticized as a "black box" which produces good results but no problem insights and explanations. Thus, how to understand and explain neural models has been an important topic in both Machine Learning and IR communities. To the best of our knowledge, the explainability of neural ranking models has not been fully studied. Instead, there have been a few papers on analyzing and understanding the empirical effect of different model components in IR tasks.

For example, Pang, Lan, Guo, Xu, and Cheng (2016a) conducted an extensive analysis on the MatchPyramid model in ad-hoc retrieval and compared different kernals, pooling sizes, and similarity functions in terms of retrieval performance. Cohen, O'Connor, and Croft (2018c) extracted the internal representations of neural ranking models and evaluated their effectiveness in four natural language processing tasks. They find that topical relevance information is usually captured in the high-level layers of a neural model. Nie, Li, and Nie (2018b) conducted empirical studies on the interaction-based neural ranking model to understand what have been learned in each neural network layer. They also notice that low-level network layers tend to capture detailed text information while high-level layers tend to have higher topical information abstraction.

While the paradigms of analyzing neural ranking models often rely on a deep understanding of specific model structure, Cohen et al. (2016) argue that there are some general patterns of which types of neural models are more suitable for each IR task. For

example, retrieval tasks with fine granularity (e.g., factoid QA) usually need higher levels of information abstraction and semantic matching, while retrieval tasks with coarse granularity (e.g., document retrieval) often rely on the exact matching or interaction between query words and document words.

Overall, the research area on the explainability of neural ranking models is largely unexplored up till now. Some skepticism about neural ranking models is also related to this, e.g., what new things can be learned by neural ranking models? It is a very challenging and promising direction for researchers in neural IR.

8. Conclusion

The purpose of this survey is to summarize the current research status on neural ranking models, analyze the existing methodologies, and gain some insights for future development. We introduced a unified formulation over the neural ranking models, and reviewed existing models based on this formulation from different dimensions under model architecture and model learning. For model architecture analysis, we reviewed existing models to understand their underlying assumptions and major design principles, including how to treat the inputs, how to consider the relevance features, and how to make evaluation. For model learning analysis, we reviewed popular learning objectives and training strategies adopted for neural ranking models. To better understand the current status of neural ranking models on major applications, we surveyed published empirical results on the ad-hoc retrieval and QA tasks to conduct a comprehensive comparison. In addition, we discussed several trending topics that are important or might be promising in the future.

Just as there has been an explosion in the development of many deep learning based methods, research on neural ranking models has increased rapidly and broadened in terms of applications. We hope this survey can help researchers who are interested in this direction, and will motivate new ideas by looking at past successes and failures. Neural ranking models are part of the broader research field of neural IR, which is a joint domain of deep learning and IR technologies with many opportunities for new research and applications. We are expecting that, through the efforts of the community, significant breakthroughs will be achieved in this domain in the near future, similar to those happened in computer vision or NLP.

Acknowlgedgments

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants no. 61425016 and 61722211, and the Youth Innovation Promotion Association CAS under Grants no. 20144310. This work was supported in part by the UMass Amherst Center for Intelligent Information Retrieval and in part by NSF IIS-1715095. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Abujabal, A., Roy, R. S., Yahya, M., & Weikum, G. (2019). *ComQA: A community-sourced dataset for complex factoid question answering with paraphrase clusters. Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL.

Ahmad, W. U., Chang, K.-W., & Wang, H. (2018). *Multi-task learning for document ranking and query suggestion. Proceedings of the sixth international conference on learning representations* ICLR

Ai, Q., Bi, K., Guo, J., & Croft, W. B. (Bi, Guo, Croft, 2018a). *Learning a deep listwise context model for ranking refinement. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval*. ACM 135–144

Ai, Q., Mao, J., Liu, Y., & Croft, W. B. (Mao, Liu, Croft, 2018b). *Unbiased learning to rank: Theory and practice. Proceedings of the 27th ACM international conference on information and knowledge management*. ACM2305–2306.

Ai, Q., Wang, X., Golbandi, N., Bendersky, M., & Najork, M. (Wang, Golbandi, Bendersky, Najork, 2018c). *Learning groupwise scoring functions using deep neural networks. WSDM'19 Workshop on Deep Matching in Practical Applications (DAPA 19)*.

Akker, B. V. D., Markov, I., & de Rijke, M. (2019). *ViTOR: Learning to rank webpages based on visual features. The World Wide Web Conference (WWW'19)*. New York, NY, USA: ACM3279–3285.

Asadi, N., Metzler, D., Elsayed, T., & Lin, J. (2011). *Pseudo test collections for learning web search ranking functions. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM1073–1082 SIGIR

Baeza-Yates, R., Ribeiro, B. D. A. N., et al. (2011). *Modern information retrieval.* New York, Harlow, England: ACM Press, Addison-Wesley.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Bennett, P. N., Radlinski, F., White, R. W., & Yilmaz, E. (2011). *Inferring and using location metadata to personalize web search. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval.* New York, NY, USA: ACM135–154 SIGIR

Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F., et al. (2012). *Modeling the impact of short- and long-term behavior on search personalization. Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval.* New York, NY, USA: ACM185–194 SIGIR

Boytsov, L., Novak, D., Malkov, Y., & Nyberg, E. (2016). *Off the beaten path: Let's replace term-based retrieval with k-NN search. Proceedings of the 25th ACM international on conference on information and knowledge management.* New York, NY, USA: ACM1099–1108 CIKM

Brenner, E., Zhao, J., Kutiyanawala, A., & Yan, Z. (2018). *End-to-end neural ranking for ecommerce product search: an application of task models and textual embeddings* arXiv:1806.07296

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., et al. (2005). *Learning to rank using gradient descent. Proceedings of the 22nd international conference on machine learning (ICML)*89–96.

Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning, 11*, 23–581.

Chapelle, O., Shivaswamy, P., Vadrevu, S., Weinberger, K., Zhang, Y., & Tseng, B. (2010). *Multi-task learning for boosting with application to web search ranking. Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM1189–1198.

Chen, H., Han, F. X., Niu, D., Liu, D., Lai, K., Wu, C., et al. (Han, Niu, Liu, Lai, Wu, et al., 2018a). *Mix: Multi-channel information crossing for text matching. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. New York, NY, USA: ACM110–119 KDD

Chen, L., Lan, Y., Pang, L., Guo, J., Xu, J., & Cheng, X. (Lan, Pang, Guo, Xu, Cheng, 2018b). *Ri-match: Integrating both representations and interactions for deep semantic matching. Information retrieval technology*. Cham: Springer International Publishing90–102.

Chen, W., Cai, F., Chen, H., & de Rijke, M. (Cai, Chen, de Rijke, 2018c). *Attention-based hierarchical neural query suggestion. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval*. New York, NY, USA: ACM1093–1096 SIGIR '18

Chen, W., Liu, T.-Y., Lan, Y., Ma, Z.-M., & Li, H. (2009). *Ranking measures and loss functions in learning to rank. Advances in neural information processing systems*315–323.

Cohen, D., Ai, Q., & Croft, W. B. (2016). *Adaptability of neural networks on varying granularity ir tasks. Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval*.
Cohen, D., Foley, J., Zamani, H., Allan, J., & Croft, W. B. (Foley, Zamani, Allan, Croft, 2018a). *Universal approximation functions for fast learning to rank: Replacing expensive regression forests with simple feed-forward networks. The 41st international ACM SIGIR conference on research & development in information retrieval*. New York, NY, USA: ACM1017–1020 SIGIR '18
Cohen, D., Mitra, B., Hofmann, K., & Croft, W. B. (Mitra, Hofmann, Croft, 2018b). *Cross domain regularization for neural ranking models using adversarial learning. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval*. ACM 1025–1028
Cohen, D., O'Connor, B., & Croft, W. B. (O'Connor, Croft, 2018c). *Understanding the representational power of neural retrieval models using NLP tasks. Proceedings of the ACM SIGIR international conference on theory of information retrieval*. ACM67–74.
Cohen, D., Yang, L., & Croft, W. B. (Yang, Croft, 2018d). *WikiPassageQA: A benchmark collection for research on non-factoid answer passage retrieval. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval, SIGIR*. Ann Arbor, MI, USA July 08–12, 2018 (pp.1165–1168
Craswell, N., Croft, W. B., Guo, J., Mitra, B., & de Rijke, M. (Croft, Guo, Mitra, de Rijke, 2017a). Report on the SIGIR 2016 workshop on neural information retrieval (Neu-IR). *50*(2), 96–103.
Craswell, N., Croft, W. B., de Rijke, M., Guo, J., & Mitra, B. (Croft, de Rijke, Guo, Mitra, 2017b). *SIGIR 2017 workshop on neural information retrieval (Neu-IR). Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM1431–1432 SIGIR '17
Dai, Z., Xiong, C., Callan, J., & Liu, Z. (2018). *Convolutional neural networks for soft-matching n-grams in ad-hoc search. Proceedings of the eleventh ACM international conference on web search and data mining*. New York, NY, USA: ACM126–134 WSDM '18
Dehghani, M., Severyn, A., Rothe, S., & Kamps, J. (Severyn, Rothe, Kamps, 2017a). Avoiding your teacher's mistakes: Training neural networks with controlled weak supervision. *CoRR, abs/1711.00313*.
Dehghani, M., Zamani, H., Severyn, A., Kamps, J., & Croft, W. B. (Zamani, Severyn, Kamps, Croft, 2017b). *Neural ranking models with weak supervision. Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. New York, NY, USA: ACM SIGIR '17
Dietz, L., Verma, M., Radlinski, F., & Craswell, N. (2017). *TREC complex answer retrieval overview. Proceedings of the twenty-sixth text retrieval conference, TREC*. Gaithersburg, Maryland, USA November 15–17, 2017
Duh, K., & Kirchhoff, K. (2008). *Learning to rank with partially-labeled data. Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM251–258 SIGIR '08
Fan, Y., Guo, J., Lan, Y., Xu, J., Pang, L., & Cheng, X. (Guo, Lan, Xu, Pang, Cheng, 2017a). *Learning visual features from snapshots for web search. Proceedings of the 2017 ACM on conference on information and knowledge management*. ACM247–256.
Fan, Y., Guo, J., Lan, Y., Xu, J., Zhai, C., & Cheng, X. (2018). *Modeling diverse relevance patterns in ad-hoc retrieval. The 41st international ACM SIGIR conference on research & development in information retrieval*. New York, NY, USA: ACM375–384 SIGIR '18
Fan, Y., Pang, L., Hou, J., Guo, J., Lan, Y., & Cheng, X. (Pang, Hou, Guo, Lan, Cheng, 2017b). *Matchzoo: A toolkit for deep text matching. CoRR* abs/1707.07270.
Feng, M., Xiang, B., Glass, M. R., Wang, L., & Zhou, B. (2015). Applying deep learning to answer selection: A study and an open task. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE,* 813–820.
Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communication of the ACM, 30*(11), 964–971.
Gao, J., Galley, M., & Li, L. (2019). *Neural approaches to conversational AI. Foundations and Trends in Information Retrieval*. Now Publisher Inc.
Ghazvininejad, M., Brockett, C., Chang, M., Dolan, B., Gao, J., Yih, W., et al. (2018). *A knowledge-grounded neural conversation model. Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI)*5110–5117.
Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies, 10*(1), 1–309.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). *Generative adversarial nets. Advances in neural information processing systems*2672–2680.
Grbovic, M., Djuric, N., Radosavljevic, V., Silvestri, F., & Bhamidipati, N. (2015). *Context- and content-aware embeddings for query rewriting in sponsored search. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM383–392 SIGIR '15
Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2016). *A deep relevance matching model for ad-hoc retrieval. Proceedings of the 25th ACM international on conference on information and knowledge management*. New York, NY, USA: ACM55–64 CIKM '16
Gysel, C. V., de Rijke, M., & Kanoulas, E. (2018). Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems, 36*(4), 38:1–38:25.
He, X., Gao, J., & Deng, L. (2014). *Deep learning for natural language processing: Theory and practice.* Redmond, WA: Deep Learning Technology Center Microsoft Research.
Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine, 29*, 82–97.
Hoogeveen, D., Verspoor, K. M., & Baldwin, T. (2015). *CQADupStack: A benchmark data set for community question-answering research. Proceedings of the 20th Australasian document computing symposium*. ACM3.
Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). *Convolutional neural network architectures for matching natural language sentences. Advances in neural information processing systems 27*. Curran Associates, Inc.2042–2050.
Huang, J., Yao, S., Lyu, C., & Ji, D. (2017). *Multi-granularity neural sentence model for measuring short text similarity. Database systems for advanced applications*. Cham: Springer International Publishing439–455.
Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). *Learning deep structured semantic models for web search using clickthrough data. Proceedings of the 22nd ACM international conference on information & knowledge management*. New York, NY, USA: ACM2333–2338 CIKM '13
Hui, K., Yates, A., Berberich, K., & de Melo, G. (Yates, Berberich, de Melo, 2017a). *PACRR: A position-aware neural ir model for relevance matching. Conference on Empirical Methods in Natural Language Processing*. ACL1060–1069.
Hui, K., Yates, A., Berberich, K., & de Melo, G. (Yates, Berberich, de Melo, 2017b). *A position-aware deep model for relevance matching in information retrieval. EMNLP'17*.
Hui, K., Yates, A., Berberich, K., & de Melo, G. (2018). *Co-PACRR: A context-aware neural IR model for ad-hoc retrieval. Proceedings of the eleventh ACM international conference on web search and data mining*. ACM279–287.
Ji, Z., Lu, Z., & Li, H. (2014). An information retrieval approach to short text conversation. *CoRR, abs/1408.6988*.
Joachims, T. (2002). *Optimizing search engines using clickthrough data. Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*. ACM133–142.
Keikha, M., Park, J. H., & Croft, W. B. (2014). *Evaluating answer passages using summarization measures. Proceedings of the 37th international ACM SIGIR conference on research &38; development in information retrieval*. New York, NY, USA: ACM963–966 SIGIR '14
Kraska, T., Beutel, A., Chi, E. H., Dean, J., & Polyzotis, N. (2018). *The case for learned index structures. Proceedings of the international conference on management of data*. New York, NY, USA: ACM489–504 SIGMOD '18
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25*. Curran Associates, Inc.1097–1105.
Lavrenko, V., & Croft, W. B. (2001). *Relevance based language models. Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*. ACM120–127.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.
Li, B., Cheng, P., & Jia, L. (Cheng, Jia, 2018a). *Joint learning from labeled and unlabeled data for information retrieval. Proceedings of the 27th international conference on computational linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics293–302 COLING '18
Li, C., Sun, Y., He, B., Wang, L., Hui, K., Yates, A., et al. (Sun, He, Wang, Hui, Yates, et al., 2018b). *NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval. Proceedings of the conference on empirical methods in natural language processing* EMNLP '18
Li, H. (2011). *Learning to rank for information retrieval and natural language processing.* Morgan & Claypool Publishers.
Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval, 3*(3), 225–331.
Liu, X., Gao, J., He, X., Deng, L., Duh, K., & Wang, Y.-Y. (2015). *Representation learning using multi-task deep neural networks for semantic classification and information retrieval. Proc. NAACL*912–921.
Liu, X., Wang, C., Leng, Y., & Zhai, C. (Wang, Leng, Zhai, 2018a). *LinkSO: A dataset for learning to retrieve similar question answer pairs on software development forums.*

Proceedings of the 4th ACM SIGSOFT international workshop on NLP for software engineering. ACM2–5.
Liu, Z., Xiong, C., Sun, M., & Liu, Z. (Xiong, Sun, Liu, 2018b). *Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. Proceedings of the 56th annual meeting of the association for computational linguistics*. Association for Computational Linguistics2395–2405.
Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). *The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*285–294.
Lu, Z., & Li, H. (2013). *A deep architecture for matching short texts. Advances in neural information processing systems 26*. Curran Associates, Inc.1367–1375.
Matthijs, N., & Radlinski, F. (2011). *Personalizing web search using long term browsing history. Proceedings of the fourth ACM international conference on web search and data mining*. New York, NY, USA: ACM25–34 WSDM'11
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (Chen, Corrado, Dean, 2013a). *Efficient estimation of word representations in vector space. In ICLR Workshop Papers*.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (Sutskever, Chen, Corrado, Dean, 2013b). *Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 26*. Curran Associates, Inc.3111–3119.
Mitra, B., & Craswell, N. (2017). Neural models for information retrieval. *CoRR* abs/1705.01509
Mitra, B., Diaz, F., & Craswell, N. (2017). *Learning to match using local and distributed representations of text for web search. Proceedings of the 26th international conference on world wide web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee1291–1299 WWW '17
Mitra, B., Simon, G., Gao, J., Craswell, N., & Deng, L. A proposal for evaluating answer distillation from web data. In Proceeding of the SIGIR 2016 WebQA Workshop.
Mollá, D., & Vicedo, J. L. (2007). Question answering in restricted domains: An overview. *Computational Linguistics, 33*(1), 41–61.
Moschitti, A., Márquez, L., Nakov, P., Agichtein, E., Clarke, C., & Szpektor, I. (2016). *SIGIR 2016 workshop webQA II: Web question answering beyond factoids. Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM1251–1252 SIGIR '16
Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., et al. (2017). *SemEval-2017 task 3: Community question answering. Proceedings of the 11th international workshop on semantic evaluation*. Vancouver, Canada: Association for Computational Linguistics SemEval '17
Nguyen, G., Tamine, L., Soulier, L., & Bricon-Souf, N. (Tamine, Soulier, Bricon-Souf, 2016a). *Toward a deep neural approach for knowledge-based IR. Workshop on Neural Information Retrieval during the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*. ACM.
Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (Rosenberg, Song, Gao, Tiwary, Majumder, Deng, 2016b). MS MARCO: A human generated machine reading comprehension dataset. *CoRR, abs/1611.09268*.
Nie, Y., Li, Y., & Nie, J. (Li, & Nie, 2018a). *Empirical study of multi-level convolution models for ir based on representations and interactions. Proceeding of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*. ACM59–66.
Nie, Y., Li, Y., & Nie, J.-Y. (Li, Nie, 2018b). *Empirical study of multi-level convolution models for ir based on representations and interactions. Proceedings of the ACM SIGIR international conference on theory of information retrieval*. ACM59–66.
Nie, Y., Sordoni, A., & Nie, J.-Y. (Sordoni, Nie, 2018c). *Multi-level abstraction convolutional model with weak supervision for information retrieval. The 41st international ACM SIGIR conference on research & development in information retrieval*. New York, NY, USA: ACM SIGIR '18
Onal, K. D., Zhang, Y., Altingovde, I. S., Rahman, M. M., Karagoz, P., Braylan, A., et al. (2018). Neural information retrieval: At the end of the early years. *Information Retrieval Journal, 21*(2–3), 111–182.
Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., et al. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language, 24*(4), 694–707.
Pang, L., Lan, Y., Guo, J., Xu, J., & Cheng, X. (Lan, Guo, Xu, Cheng, 2016a). A study of matchpyramid models on ad-hoc retrieval. *CoRR* abs/1606.04648.
Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., & Cheng, X. (Lan, Guo, Xu, Wan, Cheng, 2016b). *Text matching as image recognition. Thirtieth AAAI conference on artificial intelligence*.
Pang, L., Lan, Y., Guo, J., Xu, J., Xu, J., & Cheng, X. (2017). *DeepRank: A new deep architecture for relevance ranking in information retrieval. Proceedings of the ACM on conference on information and knowledge management*. New York, NY, USA: ACM257–266 CIKM '17
Ponte, J. M., & Croft, W. B. (1998). *A language modeling approach to information retrieval*. University of Massachusetts at Amherst Ph.D. Thesis.
Qiu, X., & Huang, X. (2015). *Convolutional neural tensor network architecture for community-based question answering. Proceedings of the 24th international conference on artificial intelligence*. AAAI Press1305–1311 IJCAI'15
Qu, C., Yang, L., Croft, W. B., Trippas, J., Zhang, Y., & Qiu, M. (2018). *Analyzing and characterizing user intent in information-seeking conversations. Proceedings of the SIGIR*.
Qu, C., Yang, L., Croft, W. B., Zhang, Y., Trippas, J., & Qiu, M. (2019). *User intent prediction in information-seeking conversations. Proceedings of the CHIIR*.
Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). *Squad: 100, 000+ questions for machine comprehension of text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*2383–2392.
Rao, J., He, H., & Lin, J. (2016). *Noise-contrastive estimation for answer selection with deep neural networks. Proceedings of the 25th ACM international on conference on information and knowledge management* CIKM '16
Rao, J., Yang, W., Zhang, Y., Ture, F., & Lin, J. J. (2019). *Multi-perspective relevance matching with hierarchical convnets for social media search* Proceedings of the national conference on artificial intelligence
Richardson, M. (2013). *MCTest: A challenge dataset for the open-domain machine comprehension of text. Proceeding of the 2013 Conference on Empirical Methods in Natural Language Processing*193–203.
Ritter, A., Cherry, C., & Dolan, W. B. (2011). *Data-driven response generation in social media. Proceedings of the conference on empirical methods in natural language processing*. Stroudsburg, PA, USA: Association for Computational Linguistics583–593 EMNLP '11
Robertson, S., & Walker, S. (1994). *Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. Proceedings of the SIGIR*.
Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129–146.
Salakhutdinov, R., & Hinton, G. (2009). Semantic hashing. *International Journal of Approximate Reasoning, 50*(7), 969–978.
Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.
Severyn, A., & Moschitti, A. (2015). *Learning to rank short text pairs with convolutional deep neural networks. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM373–382 SIGIR '15
Shang, L., & Sakai, T. (2016). *Overview of the NTCIR-12 short text conversation task. Proceeding of the NTCIR-12*473–484.
Shen, X., Tan, B., & Zhai, C. (2005). *Context-sensitive information retrieval using implicit feedback*. New York, NY, USA: ACM43–50 SIGIR '05
Shen, Y., Deng, Y., Yang, M., Li, Y., Du, N., Fan, W., et al. (2018). *Knowledge-aware attentive neural network for ranking question answer pairs. Proceedings of the he 41st international ACM SIGIR conference on research & development in information retrieval*. New York, NY, USA: ACM901–904 SIGIR '18
Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). *A latent semantic model with convolutional-pooling structure for information retrieval. Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. New York, NY, USA: ACM101–110 CIKM '14
Shtok, A., Dror, G., Maarek, Y., & Szpektor, I. (2012). *Learning from the past: Answering new questions with past answers. Proceedings of the 21st international conference on world wide web*. New York, NY, USA: ACM759–768 WWW '12
Song, X., Feng, F., Han, X., Yang, X., Liu, W., & Nie, L. (2018). *Neural compatibility modeling with attentive knowledge distillation. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval*. New York, NY, USA: ACM5–14 SIGIR '18
Szummer, M., & Yilmaz, E. (2011). *Semi-supervised learning to rank with preference regularization. Proceedings of the 20th ACM international conference on information and knowledge management*. New York, NY, USA: ACM269–278 CIKM '11
Tang, Z., & Yang, G. H. (2019). *DeepTileBars: Visualizing term distribution for neural information retrieval. AAAI'19*.
Tay, Y., Phan, M. C., Luu, A. T., & Hui, S. C. (2017). *Learning to rank question answer pairs with holographic dual LSTM architecture. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. Shinjuku, Tokyo, Japan August 7–11, 2017 (pp. 695–704
Tay, Y., Tuan, L. A., & Hui, S. C. (2018). *Hyperbolic representation learning for fast and efficient neural question answering. Proceedings of the eleventh ACM international conference on web search and data mining, WSDM*. Marina del Rey, CA, USA583–591.
Taylor, M., Guiver, J., Robertson, S., & Minka, T. (2008). *Softrank: optimizing non-smooth rank metrics. Proceedings of theWSDM*. ACM77–86.
Ustinovskiy, Y., & Serdyukov, P. (2013). *Personalization of web-search using short-term browsing context*. New York, NY, USA: ACM1979–1988 CIKM '13
Voorhees, E. M., & Tice, D. M. (2000). *Building a question answering test collection. Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM200–207 SIGIR '00
Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., et al. (2014). *Deep learning for content-based image retrieval: A comprehensive study. Proceedings of the 22nd*

ACM international conference on multimedia. New York, NY, USA: ACM157–166 MM '14
Wan, S., Lan, Y., Guo, J., Xu, J., Pang, L., & Cheng, X. (Lan, Guo, Xu, Pang, Cheng, 2016a). *A deep architecture for semantic matching with multiple positional sentence representations. Proceedings of the thirtieth AAAI conference on artificial intelligence*. AAAI Press2835–2841 AAAI'16
Wan, S., Lan, Y., Xu, J., Guo, J., Pang, L., & Cheng, X. (Lan, Xu, Guo, Pang, Cheng, 2016b). *Match-SRNN: Modeling the recursive matching structure with spatial RNN. Proceedings of the twenty-fifth international joint conference on artificial intelligence*. AAAI Press2922–2928 IJCAI'16
Wang, B., Liu, K., & Zhao, J. (2016). *Inner attention based recurrent neural networks for answer selection. Proceedings of the 54th annual meeting of the association for computational linguistics*. Association for Computational Linguistics1288–1297.
Wang, D., & Nyberg, E. (2015). *A long short-term memory model for answer sentence selection in question answering. Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing*. Association for Computational Linguistics707–712.
Wang, H., Lu, Z., Li, H., & Chen, E. (2013). *A dataset for research on short-text conversations. Proceedings of the 2013 conference on empirical methods in natural language processing*935–945.
Wang, M., Smith, N. A., & Mitamura, T. (2007). *What is the jeopardy model? a quasi-synchronous grammar for QA. Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CONLL)*.
Wang, S., & Jiang, J. (2017). *A compare-aggregate model for matching text sequences. Proceedings of the 5th international conference on learning representations* ICLR'17
Wang, Z., Hamza, W., & Florian, R. (2017). *Bilateral multi-perspective matching for natural language sentences. Proceedings of the 26th international joint conference on artificial intelligence*. AAAI Press4144–4150 IJCAI'17
Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2017). *Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. Proceedings of the ACL*.
Wu, Y., Wu, W., Xu, C., & Li, Z. (2018). *Knowledge enhanced hybrid neural network for text matching. Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI)*5586–5593.
Xia, F., Liu, T.-Y., Wang, J., Zhang, W., & Li, H. (2008). *Listwise approach to learning to rank: theory and algorithm. Proceedings of the 25th international conference on machine learning*. ACM1192–1199.
Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., & Li, H. (2010). *Context-aware ranking in web search. Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM451–458 SIGIR '10
Xiong, C., Callan, J., & Liu, T.-Y. (Callan, Liu, 2017a). *Word-entity duet representations for document ranking. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM763–772 SIGIR '17
Xiong, C., Dai, Z., Callan, J., Liu, Z., & Power, R. (Dai, Callan, Liu, Power, 2017b). *End-to-end neural ad-hoc ranking with kernel pooling. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM55–64 SIGIR '17
Xu, Z., Liu, B., Wang, B., Sun, C., & Wang, X. (2016). Incorporating loose-structured knowledge into LSTM with recall gate for conversation modeling. *CoRR, abs/1605.05110*.
Yan, R., Song, Y., & Wu, H. (Song, Wu, 2016a). *Learning to respond with deep neural networks for retrieval-based human-computer conversation system. Proceedings of the SIGIR*.
Yan, R., Song, Y., Zhou, X., & Wu, H. (Song, Zhou, Wu, 2016b). *"shall I be your chat companion?": Towards an online human-computer conversation system. Proceedings of the CIKM*.
Yan, R., Zhao, D., & E., W. (2017). *Joint learning of response ranking and next utterance suggestion in human-computer conversation system. Proceedings of the SIGIR*.
Yang, L., Ai, Q., Guo, J., & Croft, W. B. (Ai, Guo, Croft, 2016a). *aNMM: Ranking short answer texts with attention-based neural matching model. Proceedings of the 25th ACM international conference on information and knowledge management, CIKM*. Indianapolis, IN, USA October 24–28, 2016 (pp 287–296
Yang, L., Ai, Q., Spina, D., Chen, R., Pang, L., Croft, W. B., et al. (Ai, Spina, Chen, Pang, Croft, et al., 2016b). *Beyond factoid QA: effective methods for non-factoid answer sentence retrieval. Proceedings of the 38th European conference on IR research*. Padua, Italy Advances in information retrievalECIR
Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., et al. (2013). *CQARank: Jointly model topics and expertise in community question answering. Proceedings of the 22nd ACM international conference on information & knowledge management*. New York, NY, USA: ACM99–108 CIKM '13
Yang, L., Qiu, M., Qu, C., Guo, J., Zhang, Y., Croft, W. B., et al. (2018). *Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval, SIGIR 2018*. Ann Arbor, MI, USA July 08–12, 2018 (pp. 245–254)
Yang, L., Zamani, H., Zhang, Y., Guo, J., & Croft, W. B. (2017). Neural matching models for question retrieval and next question prediction in conversation. *CoRR* abs/1707.05409.
Yang, W., Zhang, H., & Lin, J. (2019). Simple applications of bert for ad hoc document retrieval. *CoRR* abs/1903.10972.
Yang, Y., Yih, S. W.-t., & Meek, C. (2015). *WikiQA: A challenge dataset for open-domain question answering*. ACL - Association for Computational Linguistics.
Yih, W., Chang, M., Meek, C., & Pastusiak, A. (2013). *Question answering using enhanced lexical semantic models. Proceedings of the 51st annual meeting of the association for computational linguistics, ACL*. Sofia, Bulgaria,: The Association for Computer Linguistics1744–1753 4–9 august
Yin, W., & Schütze, H. (2015). *MultiGranCNN: An architecture for general matching of text chunks on multiple levels of granularity. Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing*. Association for Computational Linguistics63–73.
Yu, L., Hermann, K. M., Blunsom, P., & Pulman, S. (2014). Deep learning for answer sentence selection. *CoRR, abs/1412.1632*.
Zamani, H., Bendersky, M., Wang, X., & Zhang, M. (2017). *Situational context for ranking in personal search. Proceedings of the 26th international conference on world wide web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee1531–1540 WWW '17
Zamani, H., & Croft, W. B. (2016). *Embedding-based query language models. Proceedings of the ACM international conference on the theory of information retrieval*147–156 ICTIR '16
Zamani, H., & Croft, W. B. (2017). *Relevance-based word embedding. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM505–514 SIGIR '17
Zamani, H., Croft, W. B., & Culpepper, J. S. (Croft, Culpepper, 2018a). *Neural query performance prediction using weak supervision from multiple signals. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval*. New York, NY, USA: ACM105–114 SIGIR '18
Zamani, H., Dadashkarimi, J., Shakery, A., & Croft, W. B. (2016). *Pseudo-relevance feedback based on matrix factorization. Proceedings of the 25th ACM international on conference on information and knowledge management*1483–1492 CIKM '16
Zamani, H., Dehghani, M., Croft, W. B., Learned-Miller, E., & Kamps, J. (Dehghani, Croft, Learned-Miller, Kamps, 2018b). *From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. Proceedings of the 27th ACM international conference on information and knowledge management*. New York, NY, USA: ACM497–506 CIKM '18
Zamani, H., Dehghani, M., Diaz, F., Li, H., & Craswell, N. (Dehghani, Diaz, Li, Craswell, 2018c). *SIGIR 2018 workshop on learning from limited or noisy data for information retrieval. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval*. New York, NY, USA: ACM1439–1440 SIGIR '18
Zhang, J., Liu, Y., Ma, S., & Tian, Q. (2018). *Relevance estimation with multiple information sources on search engine result pages. Proceedings of the 27th ACM international conference on information and knowledge management*. ACM 627–636
Zhang, X., He, B., & Luo, T. (2016). Training query filtering for semi-supervised learning to rank with pseudo labels. *World Wide Web, 19*(5), 833–864.
Zhao, L., & Callan, J. (2010). *Term necessity prediction. Proceedings of the 19th ACM international conference on information and knowledge management*. New York, NY, USA: ACM259–268 CIKM '10
Zheng, Y., Fan, Z., Liu, Y., Luo, C., Zhang, M., & Ma, S. (2018). *Sogou-QCL: A new dataset with click relevance label. Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval*. ACM1117–1120.
Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., et al. (2016). *Multi-view response selection for human-computer conversation. Proceedings of the EMNLP*.