

Deep Relevance Ranking Using Enhanced Document-Query Interactions

Ryan McDonald^{1,2}, Georgios-Ioannis Brokos¹ and Ion Androutsopoulos¹

¹Dept. of Informatics, Athens University of Economics and Business, Greece

²Google AI

Abstract

We explore several new models for document relevance ranking, building upon the Deep Relevance Matching Model (DRMM) of Guo et al. (2016). Unlike DRMM, which uses context-insensitive encodings of terms and query-document term interactions, we inject rich context-sensitive encodings throughout our models, inspired by PACRR’s (Hui et al., 2017) convolutional n -gram matching features, but extended in several ways including multiple views of query and document inputs. We test our models on datasets from the BIOASQ question answering challenge (Tsatsaronis et al., 2015) and TREC ROBUST 2004 (Voorhees, 2005), showing they outperform BM25-based baselines, DRMM, and PACRR.

1 Introduction

Document relevance ranking, also known as *ad-hoc retrieval* (Harman, 2005), is the task of ranking documents from a large collection using the query and the text of each document only. This contrasts with standard information retrieval (IR) systems that rely on text-based signals in conjunction with network structure (Page et al., 1999; Kleinberg, 1999) and/or user feedback (Joachims, 2002). Text-based ranking is particularly important when (i) click-logs do not exist or are small, and (ii) the network structure of the collection is non-existent or not informative for query-focused relevance. Examples include various domains in digital libraries, e.g., patents (Azzopardi et al., 2010) or scientific literature (Wu et al., 2015; Tsatsaronis et al., 2015); enterprise search (Hawking, 2004); and personal search (Chirita et al., 2005).

We investigate new deep learning architectures for document relevance ranking, focusing on *term-based interaction models*, where query terms (q -terms for brevity) are scored relative to a docu-

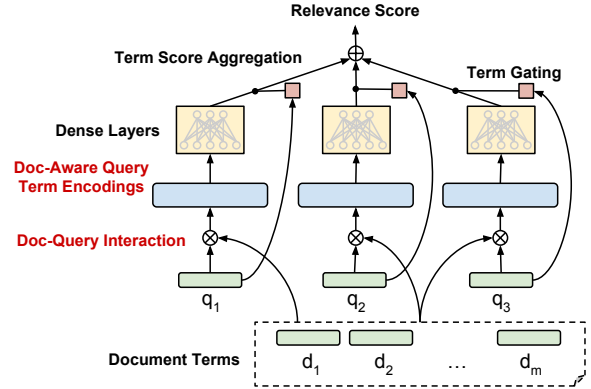


Figure 1: Illustration of DRMM (Guo et al., 2016) for a query of three terms and a document of m terms.

ment’s terms (d -terms) and their scores are aggregated to produce a relevance score for the document. Specifically, we use the Deep Relevance Matching Model (DRMM) of Guo et al. (2016) (Fig. 1), which was shown to outperform strong IR baselines and other recent deep learning methods. DRMM uses pre-trained word embeddings for q -terms and d -terms, and cosine similarity histograms (outputs of \otimes in Fig. 1), each capturing the similarity of a q -term to all the d -terms of a particular document. The histograms are fed to an MLP (dense layers of Fig. 1) that produces the (document-aware) score of each q -term. Each q -term score is then weighted using a gating mechanism (topmost box nodes in Fig. 1) that examines properties of the q -term to assess its importance for ranking (e.g., common words are less important). The sum of the weighted q -term scores is the relevance score of the document. This ignores entirely the contexts where the terms occur, in contrast to recent *position-aware* models such as PACRR (Hui et al., 2017) or those based on recurrent representations (Palangi et al., 2016).

In order to enrich DRMM with context-sensitive representations, we need to change fundamentally how q -terms are scored. This is because rich

context-sensitive representations – such as input term encodings based on RNNs or CNNs – require end-to-end training and histogram construction is not differentiable. To account for this we investigate novel query-document interaction mechanisms that are differentiable and show empirically that they are effective ways to enable end-to-end training of context-sensitive DRMM models. This is the primary contribution of this paper.

Overall, we explore several extensions to DRMM, including: PACRR-like convolutional n -gram matching features (§3.1); context-sensitive term encodings (§3.2); query-focused attention-based document representations (§3.3); pooling to reward denser term matches and turn rich term representations into fixed-width vectors (§3.4); multiple views of terms, e.g., context sensitive, insensitive, exact matches (§3.5).

We test our models on data from the BIOASQ biomedical question answering challenge (Tsatsaronis et al., 2015) and TREC ROBUST 2004 (Voorhees, 2005), showing that they outperform strong BM25-based baselines (Robertson and Zaragoza, 2009), DRMM, and PACRR.¹

2 Related Work

Document ranking has been studied since the dawn of IR; classic term-weighting schemes were designed for this problem (Sparck Jones, 1988; Robertson and Sparck Jones, 1988). With the advent of statistical NLP and statistical IR, probabilistic language and topic modeling were explored (Zhai and Lafferty, 2001; Wei and Croft, 2006), followed recently by deep learning IR methods (Lu and Li, 2013; Hu et al., 2014; Palangi et al., 2016; Guo et al., 2016; Hui et al., 2017).

Most document relevance ranking methods fall within two categories: representation-based, e.g., Palangi et al. (2016), or interaction-based, e.g., Lu and Li (2013). In the former, representations of the query and document are generated independently. Interaction between the two only happens at the final stage, where a score is generated indicating relevance. End-to-end learning and backpropagation through the network tie the two representations together. In the interaction-based paradigm, explicit encodings between pairs of queries and documents are induced. This al-

lows direct modeling of exact- or near-matching terms (e.g., synonyms), which is crucial for relevance ranking. Indeed, Guo et al. (2016) showed that the interaction-based DRMM outperforms previous representation-based methods. On the other hand, interaction-based models are less efficient, since one cannot index a document representation independently of the query. This is less important, though, when relevance ranking methods rerank the top documents returned by a conventional IR engine, which is the scenario we consider here.

One set of our experiments ranks biomedical texts. Several methods have been proposed for the BIOASQ challenge (Tsatsaronis et al., 2015), mostly based on traditional IR techniques. The most related work is of Mohan et al. (2017), who use a deep learning architecture. Unlike our work, they focus on user click data as a supervised signal, and they use context-insensitive representations of document-query interactions. The other dataset we experiment with, TREC ROBUST 2004 (Voorhees, 2005), has been used extensively to evaluate traditional and deep learning IR methods.

Document relevance ranking is also related to other NLP tasks. Passage scoring for question answering (Surdeanu et al., 2008) ranks passages by their relevance to the question; several deep networks have been proposed, e.g., Tan et al. (2015). Short-text matching/ranking is also related and has seen recent deep learning solutions (Lu and Li, 2013; Hu et al., 2014; Severyn and Moschitti, 2015). In document relevance ranking, though, documents are typically much longer than queries, which makes methods from other tasks that consider pairs of short texts not directly applicable.

Our starting point is DRMM, to which we add richer representations inspired by PACRR. Hence, we first discuss DRMM and PACRR further.

2.1 DRMM

We have already presented an overview of DRMM. For gating (topmost box nodes of Fig. 1), Guo et al. (2016) use a linear self-attention:

$$g_i = \text{softmax}\left(w_g^T \phi_g(q_i); q_1, \dots, q_n\right)$$

where $\phi_g(q_i)$ is the embedding $e(q_i)$ of the i -th q-term, or its IDF, $\text{idf}(q_i)$, and w_g is a weights vector. Gating aims to weight the (document-aware) score of each q-term (outputs of dense layers in Fig. 1) based on the importance of the term. We found that $\phi_g(q_i) = [e(q_i); \text{idf}(q_i)]$, where ‘;’ is concatenation, was optimal for all DRMM-based models.

¹The code and data of our experiments, including word embeddings, are available at <https://github.com/nlp-aueb/deep-relevance-ranking>.

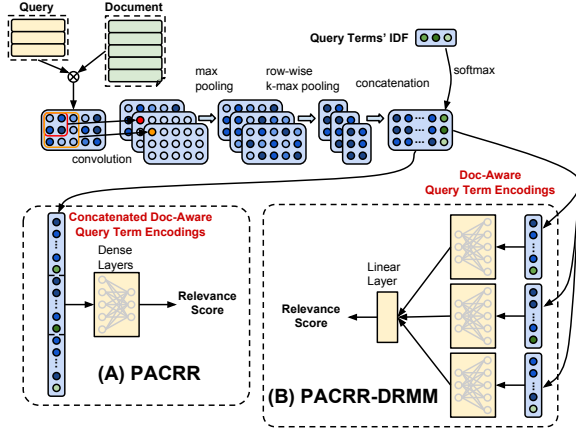


Figure 2: PACRR (Hui et al., 2017) and PACRR-DRMM.

The crux of the original DRMM are the bucketed cosine similarity histograms (outputs of \otimes nodes in Fig. 1), each capturing the similarity of a q-term to all the d-terms. In each histogram, each bucket counts the number of d-terms whose cosine similarity to the q-term is within a particular range. Consider a document with three terms, with cosine similarities, s , to a particular q-term q_i 0.5, 0.1, -0.3 , respectively. If we used two buckets $-1 \leq s < 0$ and $0 \leq s \leq 1$, then the input to the dense layers for q_i would be $\langle 1, 2 \rangle$. The fixed number of buckets leads to a fixed-dimension input for the dense layers and makes the model agnostic to different document and query lengths – one of DRMM’s main strengths. The main disadvantage is that bucketed similarities are independent of the contexts where terms occur. A q-term ‘regulated’ will have a perfect match with a d-term ‘regulated’, even if the former is ‘up regulated’ and the latter is ‘down regulated’ in context. Also, there is no reward for term matches that preserve word order, or multiple matches within a small window.

2.2 PACRR

In PACRR (Hui et al., 2017), a query-document term similarity matrix sim is computed (Fig. 2A). Each cell (i, j) of sim contains the cosine similarity between the embeddings of a q-term q_i and a d-term d_j . To keep the dimensions $l_q \times l_d$ of sim fixed across queries and documents of varying lengths, queries are padded to the maximum number of q-terms l_q , and only the first l_d terms per document are retained.² Then, convolutions (Fig. 2A) of different kernel sizes $n \times n$ ($n = 2, \dots, l_q$) are applied to sim to capture n -gram query-document similarities. For each size $n \times n$, multiple ker-

nels (filters) are used. Max pooling is then applied along the dimension of the filters (max value of all filters), followed by row-wise k -max pooling to capture the strongest k signals between each q-term and all the d-terms. The resulting matrices are concatenated into a single matrix where each row is a document-aware q-term encoding; the IDF of the q-term is also appended, normalized by applying a softmax across the IDFs of all the q-terms. Following Hui et al. (2018), we concatenate the rows of the resulting matrix into a single vector, which is passed to an MLP (Fig. 2A, dense layers) that produces a query-document relevance score.³

The primary advantage of PACRR over DRMM is that it models context via the n -gram convolutions, i.e., denser n -gram matches and matches preserving word order are encoded. However, this context-sensitivity is weak, as the convolutions operate over the similarity matrix, not directly on terms or even term embeddings. Also, unlike DRMM, PACRR requires padding and hyperparameters for maximum number of q-terms (l_q) and d-terms (l_d), since the convolutional and dense layers operate over fixed-size matrices and vectors. On the other hand, PACRR is end-to-end trainable – though Hui et al. (2017) use fixed pre-trained embeddings – unlike DRMM where the bucketed histograms are not differentiable.

3 New Relevance Ranking Models

3.1 PACRR-DRMM

In a DRMM-like version of PACRR, instead of using an MLP (dense layers, Fig. 2A) to score the *concatenation* of all the (document-aware) q-term encodings, the MLP *independently* scores each q-term encoding (the same MLP for all q-terms, Fig. 2B); the resulting scores are aggregated via a linear layer. This version, PACRR-DRMM, performs better than PACRR, using the same number of hidden layers in the MLPs. Likely this is due to the fewer parameters of its MLP, which is shared across the q-term representations and operates on shorter input vectors. Indeed, in early experiments PACRR-DRMM was less prone to over-fitting.

In PACRR-DRMM, the scores of the q-terms (outputs of dense layers, Fig. 2B) are not weighted by a gating mechanism, unlike DRMM (Fig. 1). Nevertheless, the IDFs of the q-terms, which are appended to the q-term encodings (Fig. 2B), are a

²We use PACRR-*firstk*, which Hui et al. (2017) recommend when documents fit in memory, as in our experiments.

³Hui et al. (2017) used an additional LSTM, which was later replaced by the final concatenation (Hui et al., 2018).

form of term-gating (shortcut passing on information about the terms, here their IDFs, to upper layers) applied before scoring the q-terms. By contrast, in DRMM (Fig. 1) term-gating is applied after q-term scoring, and operates on $[e(q_i); \text{idf}(q_i)]$.

3.2 Context-sensitive Term Encodings

In their original incarnations, DRMM and PACRR use pre-trained word embeddings that are insensitive to the context of a particular query or document where a term occurs. This contrasts with the plethora of systems that use context-sensitive word encodings (for each particular occurrence of a word) in virtually all NLP tasks (Bahdanau et al., 2014; Plank et al., 2016; Lample et al., 2016). In general, this is achieved via RNNs, e.g., LSTMs (Gers et al., 2000), or CNNs (Bai et al., 2018).

In the IR literature, context-sensitivity is typically viewed through two lenses: term proximity (Büttcher et al., 2006) and term dependency (Metzler and Croft, 2005). The former assumes that the context around a term match is also relevant, whereas the latter aims to capture when multiple terms (e.g., an n -gram) must be matched together. An advantage of neural network architectures like RNNs and CNNs is that they can capture both.

In the models below (§§3.3–3.4), an encoder produces the *context-sensitive encoding* of each q-term or d-term from the pre-trained embeddings. To compute this we use a standard BiLSTM encoding scheme and set the context-sentence encoding as the concatenation of the last layer’s hidden states of the forward and backward LSTMs at each position. As is common for CNNs and even recent RNN term encodings (Peters et al., 2018), we use the original term embedding $e(t_i)$ as a residual and combine it with the BiLSTM encodings. Specifically, if $\vec{h}(t_i)$ and $\overleftarrow{h}(t_i)$ are the last layer’s hidden states of the left-to-right and right-to-left LSTMs for term t_i , respectively, then we set the context-sensitive term encoding as:

$$c(t_i) = [\vec{h}(t_i) + e(t_i); \overleftarrow{h}(t_i) + e(t_i)] \quad (1)$$

Since we are adding the original term embedding to each LSTM hidden state, we require the dimensionality of the hidden layers to be equal to that of the original embedding. Other methods were tried, including passing all representations through an MLP, but these had no effect on performance.

This is an orthogonal way to incorporate context into the model relative to PACRR. PACRR creates

a query-document similarity matrix and computes n -gram convolutions over the matrix. Here we incorporate context directly into the term encodings; hence similarities in this space are already context-sensitive. One way to view this difference is the point at which context enters the model – directly during term encoding (Eq. 1) or after term similarity scores have been computed (PACRR, Fig. 2).

3.3 ABEL-DRMM

Using the context-sensitive q-term and d-term encodings of §3.2 (Eq. 1), our next extension to DRMM is to create *document-aware* q-term encodings that go beyond bucketed histograms of cosine similarities, the stage in Fig. 1 indicated by \otimes . We focus on differentiable encodings to facilitate end-to-end training from inputs to relevance scores.

Figure 3 shows the sub-network that computes the document-aware encoding of a q-term q_i in the new model, given a document $d = \langle d_1, \dots, d_m \rangle$ of m d-terms. We first compute a dot-product⁴ attention score $a_{i,j}$ for each d_j relative to q_i :

$$a_{i,j} = \text{softmax}(c(q_i)^T c(d_j); d_1, \dots, d_m) \quad (2)$$

where $c(t)$ is the context-sensitive encoding of t (Eq. 1). We then sum the context-sensitive encodings of the d-terms, weighted by their attention scores, to produce an attention-based representation d_{q_i} of document d from the viewpoint of q_i :

$$d_{q_i} = \sum_j a_{i,j} c(d_j) \quad (3)$$

The Hadamard product (element-wise multiplication, \odot) between the (L2-normalized) document representation d_{q_i} and the q-term encoding $c(q_i)$ is then computed and used as the fixed-dimension document-aware encoding $\phi_H(q_i)$ of q_i (Fig. 3):

$$\phi_H(q_i) = \frac{d_{q_i}}{\|d_{q_i}\|} \odot \frac{c(q_i)}{\|c(q_i)\|} \quad (4)$$

The \otimes nodes and lower parts of the DRMM network of Fig. 1 are now replaced by (multiple copies of) the sub-network of Fig. 3 (one copy per q-term), with the \odot nodes replacing the \otimes nodes. We call the resulting model Attention-Based Element-wise DRMM (ABEL-DRMM).

Intuitively, if the document contains one or more terms d_j that are similar to q_i , the attention

⁴Dot-products have a larger range than other similarity functions, encouraging low entropy attention distributions.

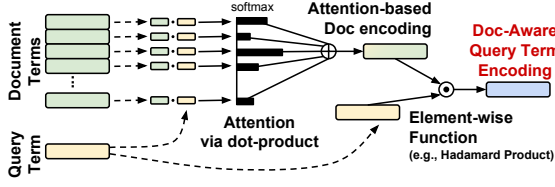


Figure 3: ABEL-DRMM sub-net. From context-aware q-term and d-term encodings (Eq. 1), it generates fixed-dimension *document-aware q-term encodings* to be used in DRMM (Fig. 1, replacing \otimes nodes).

mechanism will have emphasized mostly those terms and, hence, d_{q_i} will be similar to $c(q_i)$, otherwise not. This similarity could have been measured by the cosine similarity between d_{q_i} and $c(q_i)$, but the cosine similarity assigns the same weight to all the dimensions, i.e., to all the (L2 normalized) element-wise products in $\phi_H(q_i)$, which cosine similarity just sums. By using the Hadamard product, we pass on to the upper layers of DRMM (the dense layers of Fig. 1), which score each q-term with respect to the document, all the (normalized) element-wise products of $\phi_H(q_i)$, allowing the upper layers to learn which element-wise products (or combinations of them) are important when matching a q-term to the document.

Other element-wise functions can also be used to compare d_{q_i} to $c(q_i)$, instead of the Hadamard product (Eq. 4). For example, a vector containing the squared terms of the Euclidean distance between d_{q_i} and $c(q_i)$ could be used instead of $\phi_H(q_i)$. This change had no effect on ABEL-DRMM’s performance on development data. We also tried using $[d_{q_i}; c(q_i)]$ instead of $\phi_H(q_i)$, but performance on development data deteriorated.

ABEL-DRMM is agnostic to document length, like DRMM. ABEL-DRMM, however, is trainable end-to-end, unlike the original DRMM. Still, both models do not reward higher density matches.

3.4 POSIT-DRMM

Ideally, we want models to reward both the maximum match between a q-term and a document, but also the average match (between several q-terms and the document) to reward documents that have a higher density of matches. The document-aware q-term scoring of ABEL-DRMM does not account for this, as the attention summation hides whether a single or multiple terms were matched with high similarity. We also want models to be end-to-end trainable, like ABEL-DRMM.

Figure 4 (context-sensitive box) outlines a simple network that produces document-aware q-

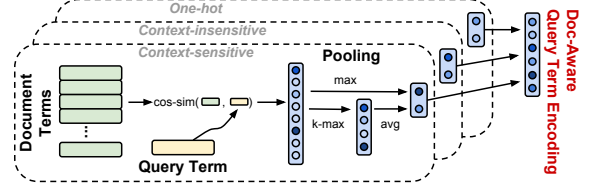


Figure 4: POSIT-DRMM with multiple views (+MV). Three two-dimensional document-aware q-term encodings, one from each view, are produced, concatenated, and used in DRMM (Fig. 1, replacing \otimes nodes).

term encodings, replacing the ABEL-DRMM sub-network of Fig. 3 in the DRMM framework. We call the resulting model POoled SIMilarity DRMM (POSIT-DRMM). As in ABEL-DRMM, we compute an attention score $a_{i,j}$ for each d_j relative to q_i , now using cosine similarity (cf. Eq. 2):

$$a_{i,j} = \frac{c(q_i)^T c(d_j)}{\|c(q_i)\| \|c(d_j)\|} \quad (5)$$

However, we do not use the $a_{i,j}$ scores to compute a weighted average of the encodings of the d-terms (cf. Eq. 3), which is also why there is no softmax in $a_{i,j}$ above (cf. Eq. 2).⁵ Instead, we concatenate the attention scores of the m d-terms:

$$a_i = \langle a_{i,1}, \dots, a_{i,j}, \dots, a_{i,m} \rangle^T$$

and we apply two pooling steps on a_i to create a 2-dimensional document-aware encoding $\phi_P(q_i)$ of the q-term q_i (Fig. 4). First max-pooling, which returns the single best match of q_i in the document. Then average pooling over a k -max-pooled version of a_i , which represents the average similarity for the top k matching terms:

$$\phi_P(q_i) = \langle \max(a_i), \text{avg}(k\text{-max}(a_i)) \rangle^T$$

POSIT-DRMM has many fewer parameters than the other models. The input to the upper q-term scoring dense layers of the DRMM framework (Fig. 1) for ABEL-DRMM has the same dimensionality as pre-trained term embeddings, on the order of hundreds. By contrast, the input dimensionality here is 2. Hence, POSIT-DRMM does not require deep dense layers, but uses a single layer (depth 1). More information on hyperparameters is provided in Appendix A (supplementary material).

POSIT-DRMM is closely related to PACRR (and PACRR-DRMM). Like POSIT-DRMM, PACRR first

⁵The $a_{i,j}$ s still need to be normalized for input to the upper layers, but they do not need to be positive summing to 1. This is why we use cosine similarity in Eq. 5 instead of dot-products combined with softmax of Eq. 2.

computes cosine similarities between all q-terms and d-terms (Fig. 2). It then applies n -gram convolutions to the similarity matrix to inject context-awareness, and then pooling to create document-aware q-term representations. Instead, POSIT-DRMM relies on the fact that the term encodings are now already context sensitive (Eq. 1) and thus skips the n -gram convolutions. Again, this is a choice of when context is injected – during term encoding or after computing similarity scores.

Mohan et al.’s work (2017) is related in the sense that for each q-term, document-aware encodings are built over the best matching (Euclidean distance) d-term. But again, term encodings are context-insensitive pre-trained word embeddings and the model is not trained end-to-end.

3.5 Multiple Views of Terms (+MV)

An extension to ABEL-DRMM and POSIT-DRMM (or any deep model) is to use multiple views of terms. The basic POSIT-DRMM produces a two-dimensional document-aware encoding of each q-term (Fig. 4, context-sensitive box) viewing the terms as their context-sensitive encodings (Eq. 1). Another two-dimensional document-aware q-term encoding can be produced by viewing the terms directly as their pre-trained embeddings without converting them to context-sensitive encodings (Fig. 4, context-insensitive box). A third view uses one-hot vector representations of terms, which allows exact term matches to be modeled, as opposed to near matches in embedding space. Concatenating the outputs of the 3 views, we obtain 6-dimensional document-aware q-term encodings, leading to a model dubbed POSIT-DRMM+MV. An example of this multi-view document-aware query term representation is given in Fig. 5 for a query-document pair from BIOASQ’s development data.

The multi-view extension of ABEL-DRMM (ABEL-DRMM+MV) is very similar, i.e., it uses context-sensitive term encodings, pre-trained term embeddings, and one-hot term encodings in its three views. The resulting three document-aware q-term embeddings can be summed or concatenated, though we found the former more effective.

3.6 Alternative Network Structures

The new models (§§3.1–3.5) were selected by experimenting on development data. Many other extensions were considered, but not ultimately used as they were not beneficial empirically, including deeper and wider RNNs or CNN encoders (Bai

et al., 2018); combining document-aware encodings from all models; different attention mechanisms, e.g., multi-head (Vaswani et al., 2017).

Pointer Networks (Vinyals et al., 2015) use the attention scores directly to select an input component. POSIT-DRMM does this via max and average pooling, not argmax. We implemented Pointer Networks – argmax over ABEL-DRMM attention to select the best d-term encoding – but empirically this was similar to ABEL-DRMM. Other architectures considered in the literature include the K-NRM model of Xiong et al. (2017). This is similar to both ABEL-DRMM and POSIT-DRMM in that it can be viewed as an end-to-end version of DRMM. However, it uses kernels over the query-document interaction matrix to produce features per q-term.

The work of Pang et al. (2017) is highly related and investigates many different structures, specifically aimed at incorporating context-sensitivity. However, unlike our work, Pang et al. first extract contexts (n -grams) of documents that match q-terms. Multiple interaction matrices are then constructed for the entire query relative to each of these contexts. These document contexts may match one or more q-terms allowing the model to incorporate term proximity. These interaction matrices can also be constructed using exact string match similar to POSIT-DRMM+MV.

4 Experiments

We experiment with ad-hoc retrieval datasets with hundreds of thousands or millions of documents. As deep learning models are computationally expensive, we first run a traditional IR system⁶ using the BM25 score (Robertson and Zaragoza, 2009) and then re-rank the top N returned documents.

4.1 Methods Compared

All systems use an extension proposed by Severyn and Moschitti (2015), where the relevance score is combined via a linear model with a set of *extra features*. We use four extra features: z-score normalized BM25 score; percentage of q-terms with exact match in the document (regular and IDF weighted); and percentage of q-term bigrams matched in the document. The latter three features were taken from Mohan et al. (2017).

In addition to the models of §§2.1, 2.2, 3.1–3.5, we used the following baselines: Standard Okapi

⁶We used Galago (<http://www.lemurproject.org/galago.php>, v.3.10). We removed stop words and applied Krovetz’s stemmer (Krovetz, 1993).

BM25 (**BM25**); and BM25 re-ranked with a linear model over the four extra features (**BM25+extra**). These IR baselines are very strong and most recently proposed deep learning models do *not* beat them.⁷ DRMM and PACRR are also strong baselines and have shown superior performance over other deep learning models on a variety of data (Guo et al., 2016; Hui et al., 2017).⁸

All hyperparameters were tuned on development data and are available in Appendix A. All models were trained using Adam (Kingma and Ba, 2014) with batches containing a randomly sampled negative example per positive example⁹ and a pair-wise loss. As the datasets contain only documents marked as relevant, negative examples were sampled from the top N documents (returned by BM25) that had not been marked as relevant.

We evaluated the models using the TREC ad-hoc retrieval evaluation script¹⁰ focusing on MAP, Precision@20 and nDCG@20 (Manning et al., 2008). We trained each model five times with different random seeds and report the mean and standard deviation for each metric on test data; in each run, the model selected had the highest MAP on the development data. We also report results for an **oracle**, which re-ranks the N documents returned by BM25 placing all human-annotated relevant documents at the top. To test for statistical significance between two systems, we employed two-tailed stratified shuffling (Smucker et al., 2007; Dror et al., 2018) using the model with the highest development MAP over the five runs per method.

4.2 BioASQ Experiments

Our first experiment used the dataset of the document ranking task of BIOASQ (Tsatsaronis et al., 2015), years 1–5.¹¹ It contains 2,251 English biomedical questions, each formulated by a biomedical expert, who searched (via PubMed¹²) for, and annotated relevant documents. Not all relevant documents were necessarily annotated, but the data includes additional expert relevance judg-

ments made during the official evaluation.¹³

The document collection consists of approx. 28M ‘articles’ (titles and abstracts only) from the ‘MEDLINE/PubMed Baseline 2018’ collection.¹⁴ We discarded the approx. 10M articles that contained only titles, since very few of these were annotated as relevant. For the remaining 18M articles, a document was the concatenation of each title and abstract. Consult Appendix B for further statistics of the dataset. Word embeddings were pre-trained by applying word2vec (Mikolov et al., 2013) (see Appendix A for hyper-parameters) to the 28M ‘articles’ of the MEDLINE/PubMed collection. IDF values were computed over the 18M articles that contained both titles and abstracts.

The 1,751 queries of years 1–4 were used for training, the first 100 queries of year 5 (batch 1) for development, and the remaining 400 queries of year 5 (batches 2–5) as test set. We set $N = 100$, since even using only the top 100 documents of BM25, the oracle scores are high. PubMed articles published after 2015 for the training set, and after 2016 for the development and test sets, were removed from the top N (and replaced by lower ranked documents up to N), as these were not available at the time of the human annotation.

Table 1 reports results on the BIOASQ test set, averaged over five runs as well as the single best run (by development MAP) with statistical significance. The enhanced models of this paper perform better than BM25 (even with extra features), PACRR, and DRMM. There is hardly any difference between PACRR and DRMM, but our combination of the two (PACRR-DRMM) surpasses them both on average, though the difference is statistically significant ($p < 0.05$) only when comparing to PACRR. Models that use context-sensitive term encodings (ABEL-DRMM, POSIT-DRMM) outperform other models, even PACRR-style models that incorporate context at later stages in the network. This is true both on average and by statistical significance over the best run. The best model on average is POSIT-DRMM+MV, though it is not significantly different than POSIT-DRMM.

4.3 TREC Robust 2004 Experiments

Our primary experiments were on the BIOASQ dataset as it has one of the largest sets of queries

⁷See, for example, Table 2 of Guo et al. (2016).

⁸For PACRR/PACRR-DRMM, we used/modified the code released by Hui et al. (2017, 2018). We use our own implementation of DRMM, which performs roughly the same as Guo et al. (2016), though the results are not directly comparable due to different random partitions of the data.

⁹We limit positive examples to be in the top N documents.

¹⁰https://trec.nist.gov/trec_eval/ (v9.0)

¹¹<http://bioasq.org/>.

¹²<https://www.ncbi.nlm.nih.gov/pubmed/>

¹³Our results are, thus, not comparable to those of participating systems, since experts did not consider our outputs.

¹⁴See https://www.nlm.nih.gov/databases/download/pubmed_medline.html.

AVERAGE OVER FIVE RUNS WITH STD. DEV.			
System	MAP	P@20	nDCG@20
Traditional IR Baselines			
BM25	46.1 \pm .0	25.5 \pm .0	55.4 \pm .0
BM25+extra	48.7 \pm .0	26.6 \pm .0	58.1 \pm .0
Deep Learning Baselines			
PACRR	49.1 \pm .2	27.1 \pm .1	58.5 \pm .2
DRMM	49.3 \pm .2	27.2 \pm .2	58.5 \pm .3
Deep Learning with Enhanced Interactions			
PACRR-DRMM	49.9 \pm .1	27.4 \pm .1	59.3 \pm .1
ABEL-DRMM	50.3 \pm .2	27.5 \pm .1	59.6 \pm .2
+MV	50.4 \pm .2	27.4 \pm .2	59.7 \pm .3
POSIT-DRMM	50.7 \pm .2	27.8 \pm .1	60.1 \pm .2
+MV	51.0 \pm .1	27.9 \pm .1	60.3 \pm .2
Oracle	72.8 \pm .0	37.5 \pm .0	80.7 \pm .0

BEST RUN WITH STAT. SIG.			
System	MAP	P@20	nDCG@20
Traditional IR Baselines			
BM25	46.1	25.5	55.4
BM25+extra	48.7 (1)	26.6 (1)	58.1 (1)
Deep Learning Baselines			
PACRR	49.1 (1)	27.0 (1-2)	58.6 (1)
DRMM	49.3 (1-2)	27.1 (1-2)	58.8 (1-2)
Deep Learning with Enhanced Interactions			
PACRR-DRMM	50.0 (1-3)	27.3 (1-2)	59.4 (1-3)
ABEL-DRMM	50.2 (1-4)	27.5 (1-4)	59.4 (1-3)
+MV	50.5 (1-4)	27.6 (1-4)	59.8 (1-4)
POSIT-DRMM	50.7 (1-4,6)	27.9 (1-7)	60.1 (1-4,6)
+MV	51.0 (1-7)	27.7 (1-4)	60.3 (1-7)
Oracle	72.8	37.5	80.7

Table 1: Performance on BIOASQ test data. Statistically significant ($p < 0.05$) difference from BM25¹; BM25+extra²; PACRR³; DRMM⁴; PACRR-DRMM⁵; ABEL-DRMM⁶; ABEL-DRMM+MV⁷.

AVERAGE OVER FIVE RUNS WITH STD. DEV.			
System	MAP	P@20	nDCG@20
Traditional IR Baselines			
BM25	23.8 \pm .0	35.4 \pm .0	42.5 \pm .0
BM25+extra	25.0 \pm .0	36.7 \pm .0	43.2 \pm .0
Deep Learning Baselines			
PACRR	25.8 \pm .2	37.2 \pm .4	44.3 \pm .4
DRMM	25.6 \pm .6	37.0 \pm .8	44.4 \pm .6
Deep Learning with Enhanced Interactions			
PACRR-DRMM	25.9 \pm .4	37.3 \pm .7	44.4 \pm .7
ABEL-DRMM	26.3 \pm .4	38.0 \pm .6	45.6 \pm .4
+MV	26.5 \pm .4	38.0 \pm .5	45.5 \pm .4
POSIT-DRMM	27.0 \pm .4	38.3 \pm .6	45.7 \pm .5
+MV	27.2 \pm .3	38.6 \pm .6	46.1 \pm .4
Oracle	68.0 \pm .0	82.1 \pm .0	93.1 \pm .0

BEST RUN WITH STAT. SIG.			
System	MAP	P@20	nDCG@20
Traditional IR Baselines			
BM25	23.8	35.4	42.5
BM25+extra	25.0 (1)	36.7 (1)	43.2 (1)
Deep Learning Baselines			
PACRR	25.8 (1-2)	37.4 (1-2)	44.5 (1-2)
DRMM	25.9 (1-2)	37.2 (1-2)	44.4 (1-2)
Deep Learning with Enhanced Interactions			
PACRR-DRMM	25.9 (1-2)	37.6 (1-2,4)	44.5 (1-2)
ABEL-DRMM	26.1 (1-5)	38.0 (1-5)	45.4 (1-5)
+MV	26.4 (1-6)	38.2 (1-5)	45.8 (1-6)
POSIT-DRMM	27.1 (1-7)	38.8 (1-7)	46.2 (1-7)
+MV	27.1 (1-7)	38.9 (1-7)	46.4 (1-7)
Oracle	68.0	82.1	93.1

Table 2: Performance on TREC ROBUST test data. Statistically significant ($p < 0.05$) difference from BM25¹; BM25+extra²; PACRR³; DRMM⁴; PACRR-DRMM⁵; ABEL-DRMM⁶; ABEL-DRMM+MV⁷.

(with manually constructed relevance judgments) and document collections, making it a particularly realistic dataset. However, in order to ground our models in past work we also ran experiments on TREC ROBUST 2004 (Voorhees, 2005), which is a common benchmark. It contains 250 queries¹⁵ and 528K documents. As this dataset is quite small, we used a 5-fold cross-validation. In each fold, approx. $\frac{3}{5}$ of the queries were used for training, $\frac{1}{5}$ for development, and $\frac{1}{5}$ for testing. We applied word2vec to the 528K documents to obtain pre-trained embeddings. IDF values were computed over the same corpus. Here we used $N = 1000$, as the oracle scores for $N = 100$ were low.

Table 2 shows the TREC ROBUST results, which largely mirror those of BIOASQ. POSIT-DRMM+MV is still the best model, though again not significantly different than POSIT-DRMM. Furthermore, ABEL-DRMM and POSIT-DRMM are clearly better than the deep learning baselines,¹⁶

but unlike BIOASQ, there is no statistically significant difference between PACRR-DRMM and the two deep learning baselines. Even though the scores are quite close (particularly MAP) both ABEL-DRMM and POSIT-DRMM are statistically different from PACRR-DRMM, which was not the case for BIOASQ. ABEL-DRMM+MV is significantly different than ABEL-DRMM on the best run for MAP and nDCG@20, unlike BIOASQ where there was no statistically significant difference between the two methods. However, on average over 5 runs, the systems show little difference.

5 Discussion

An interesting question is how well the deep models do without the extra features. For BIOASQ, the best model’s (POSIT-DRMM+MV) MAP score drops from 48.1 to 46.2 on the development set, which is higher than the BM25 baseline (43.7), but on-par with BM25+EXTRA (46.0). We should

¹⁵We used the ‘title’ fields of the queries.

¹⁶The results we report for our implementation of DRMM are slightly different than those of Guo et al. (2016). There

are a number of reasons for why this might be the case: there is no standard split of the data; non-standard preprocessing of the documents; the original DRMM paper reranks the top documents returned by Query Likelihood and not BM25.

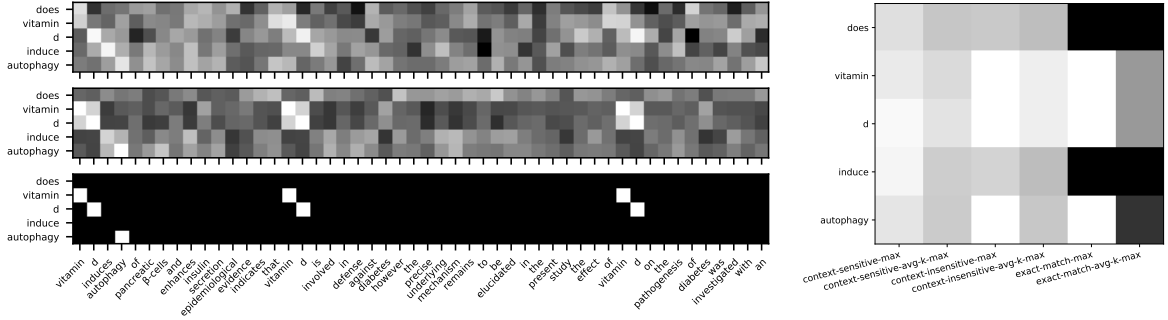


Figure 5: Left: Cosine similarities (POSIT-DRMM attention) of query and document terms, with context-sensitive, context-insensitive, and exact match views of the terms (top to bottom). Document truncated to 50 words. White is stronger. Right: Corresponding POSIT-DRMM+MV 6-dimensional document-aware query term encodings.

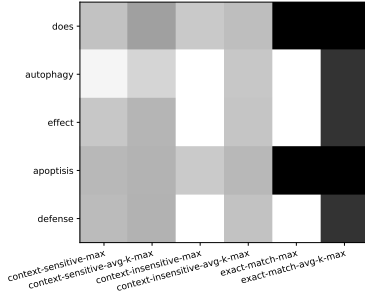


Figure 6: POSIT-DRMM+MV 6-dimensional document-aware q-term encodings for ‘Does autophagy induce apoptosis defense?’ and the same document as Fig. 5.

note, however, that on this set, the DRMM baseline without the extra features (which include BM25) is actually lower than BM25 (MAP 42.5), though it is obviously adding a useful signal, since DRMM with the extra features performs better (46.5).

We also tested the contribution of context-sensitive term encodings (Eq. 1). Without them, i.e., using directly the pre-trained embeddings, MAP on BIOASQ development data dropped from 47.6 to 46.3, and from 48.1 to 47.0 for ABEL-DRMM and POSIT-DRMM, respectively.

Fig. 5 shows the cosine similarities (attention scores, Eq. 5) between q-terms and d-terms, using term encodings of the three views (Fig. 4), for a query “Does Vitamin D induce autophagy?” and a relevant document from the BIOASQ development data. POSIT-DRMM indeed marks this as relevant. In the similarities of the context-insensitive view (middle left box) we see multiple matches around ‘vitamin d’ and ‘induce autophagy’. The former is an exact match (white squares in lower left box) and the latter a soft match. The context-sensitive view (upper left box) smooths things out and one can see a straight diagonal white line matching ‘vitamin d induce autophagy’. The right box of Fig. 5 shows the 6 components (Fig. 4) of the document-aware q-term encodings. Although

some terms are not matched exactly, the context sensitive max and average pooled components (two left-most columns) are high for all q-terms. Interestingly, ‘induce’ and ‘induces’ are not an exact match (leading to black cells for ‘induce’ in the two right-most columns) and the corresponding context-insensitive component of (third cell from left) is low. However, the two components of the context-sensitive view (two left-most cells of ‘induce’) are high, esp. the max-pooling component (left-most). Finally, ‘vitamin d’ has multiple matches leading to a high average k-max pooled value, which indicates that the importance of that phrase in the document.

Fig. 6 shows the 6 components of the document-aware q-term encodings for another query and the same document, which is now irrelevant. In the max pooling columns of the exact match and context-insensitive view (columns 3, 5), the values look quite similar to those of Fig. 5. However, POSIT-DRMM scores this query-document pair low for two reasons. First, in the average-k-max pooling columns (columns 2, 4, 6) we get lower values than Fig. 5, indicating that there is less support for this pair in terms of density. Second, the context sensitive values (columns 1, 2) are much worse, indicating that even though many exact matches exist, in context, the meaning is not the same.

We conclude by noting there is still quite a large gap between the current best models and the oracle re-ranking scores. Thus, there is head room for improvements through more data or better models.

Acknowledgements

We thank the reviewers for their constructive feedback that greatly improved this work. Oscar Täckström gave thorough input on an early draft of this work. Finally, AUEB’s NLP group provided many suggestions over the course of the work.

References

- Leif Azzopardi, Wim Vanderbauwhede, and Hideo Joho. 2010. Search system requirements of patent analysts. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 775–776, Geneva, Switzerland.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.
- Stefan Büttcher, Charles LA Clarke, and Brad Lushman. 2006. Term proximity scoring for ad-hoc retrieval on very large text collections. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 621–622. ACM.
- Paul Alexandru Chirita, Rita Gavriloiu, Stefania Ghita, Wolfgang Nejdl, and Raluca Paiu. 2005. Activity based metadata for semantic desktop search. In *European Semantic Web Conference*, pages 439–454, Heraklion, Greece.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia.
- Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471.
- Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64, Indianapolis, IN.
- Donna K Harman. 2005. The TREC ad hoc experiments. Technical report, National Institute of Standards and Technology (NIST).
- David Hawking. 2004. Challenges in enterprise search. In *Proceedings of the 15th Australasian Database Conference-Volume 27*, pages 15–24, Dunedin, New Zealand.
- Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems 27*, pages 2042–2050.
- Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A position-aware neural IR model for relevance matching. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1049–1058, Copenhagen, Denmark.
- Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2018. Co-PACRR: A context-aware neural IR model for ad-hoc retrieval. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 279–287, Marina Del Rey, CA.
- Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, Edmonton, Canada.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.
- Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202, Pittsburgh, PA.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, CA.
- Zhengdong Lu and Hang Li. 2013. A deep architecture for matching short texts. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, pages 1367–1375, Lake Tahoe, NV.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- Donald Metzler and W Bruce Croft. 2005. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119, Lake Tahoe, Nevada.

- Sunil Mohan, Nicolas Fiorini, Sun Kim, and Zhiyong Lu. 2017. Deep learning for biomedical information retrieval: Learning textual relevance from click logs. In *BioNLP 2017*, pages 222–231.
- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.
- Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):694–707.
- Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. 2017. Deeprank: A new deep architecture for relevance ranking in information retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 257–266. ACM.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.
- Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany.
- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Stephen E. Robertson and Karen Sparck Jones. 1988. Document retrieval systems. chapter Relevance Weighting of Search Terms, pages 143–160. Taylor Graham Publishing, London, UK.
- Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 373–382, Santiago, Chile.
- Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632. ACM.
- Karen Sparck Jones. 1988. Document retrieval systems. chapter A Statistical Interpretation of Term Specificity and Its Application in Retrieval, pages 132–142. Taylor Graham Publishing, London, UK.
- Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *Proceedings of the 46th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 719–727, Columbus, OH.
- Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. LSTM-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*.
- George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008, Long Beach, CA.
- Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28*, pages 2692–2700, Montreal, Canada.
- Ellen M Voorhees. 2005. The TREC robust retrieval track. *ACM SIGIR Forum*, 39(1):11–20.
- Xing Wei and W Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, Seattle, WA.
- Jian Wu, Kyle Mark Williams, Hung-Hsuan Chen, Madian Khabsa, Cornelia Caragea, Suppawong Tuarob, Alexander G Ororbia, Douglas Jordan, Prasenjit Mitra, and C Lee Giles. 2015. Citeseerx: AI in a digital library search engine. *AI Magazine*, 36(3):35–48.
- Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–64. ACM.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, New Orleans, LA.