Ranking Documents by Answer-Passage Quality

Evi Yulianti[*]
RMIT University
Melbourne, Australia

Ruey-Cheng Chen[†]
SEEK Ltd.
Melbourne, Australia

Falk Scholer
RMIT University
Melbourne, Australia

W. Bruce Croft
RMIT University
Melbourne, Australia

Mark Sanderson
RMIT University
Melbourne, Australia

ABSTRACT

Evidence derived from passages that closely represent likely answers to a posed query can be useful input to the ranking process. Based on a novel use of Community Question Answering data, we present an approach for the creation of such passages. A general framework for extracting answer passages and estimating their quality is proposed, and this evidence is integrated into ranking models. Our experiments on two web collections show that such quality estimates from answer passages provide a strong indication of document relevance and compare favorably to previous passage-based methods. Combining such evidence can significantly improve over a set of state-of-the-art ranking models, including Quality-Biased Ranking, External Expansion, and a combination of both. A final ranking model that incorporates all quality estimates achieves further improvements on both collections.

KEYWORDS

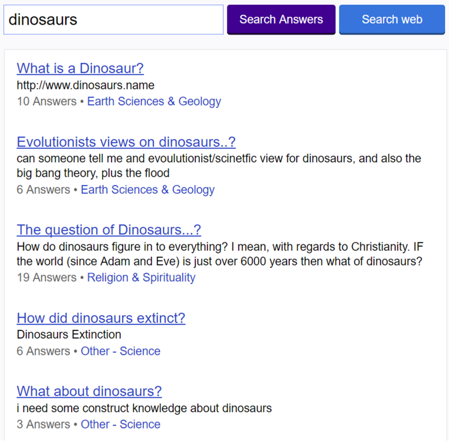
Document ranking; quality estimation; answer passages

ACM Reference Format:
Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2018. Ranking Documents by Answer-Passage Quality. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210028

1 INTRODUCTION

It has long been thought that combining document-level and passage-level evidence is an effective retrieval approach [8, 46]. Bendersky and Kurland [4], for example, showed that combining evidence from the best-matching passage in retrieved documents leads to increased retrieval effectiveness.

Different types of passages have been examined. Tombros and Sanderson [43] proposed so-called query biased summaries for use in search result pages. Later work provided evidence supporting the use of summaries as a passage representation to improve ad hoc retrieval [14, 22, 40]. Such summaries are created based on the degree of query-term matching, rather than document relevance. It remains to be seen if more effective passages can be found.



What is a Dinosaur?
http://www.dinosaurs.name
10 Answers • Earth Sciences & Geology

Evolutionists views on dinosaurs..?
can someone tell me and evoulutionist/scinetfic view for dinosaurs, and also the big bang theory, plus the flood
6 Answers • Earth Sciences & Geology

The question of Dinosaurs...?
How do dinosaurs figure in to everything? I mean, with regards to Christianity. IF the world (since Adam and Eve) is just over 6000 years then what of dinosaurs?
19 Answers • Religion & Spirituality

How did dinosaurs extinct?
Dinosaurs Extinction
6 Answers • Other - Science

What about dinosaurs?
i need some construct knowledge about dinosaurs
3 Answers • Other - Science

Figure 1: An example of questions from the CQA site, *Yahoo! Answers*, that are related to the given query "dinosaurs"

We investigate whether passages can be biased towards selecting text fragments that are more likely to bear answers to the query, and whether this new approach would give a better indication of underlying document relevance. The induced representation would tend to cover a richer set of text evidence rather than just the given query terms. We call these fragments *answer passages.*

We create answer passages by exploiting content in a specialized resource where high quality, human-curated question-answer structures are abundant: Community Question Answering (CQA) services. The text content on such services is utilized in a specific way: not to reuse or synthesize answers, but to provide an indication as to which text fragments in a document are likely to be part of an accepted answer. This "answer-bearingness" property can serve as a valuable document ranking signal.

While exploiting information from external resources to improve ranking is common [5, 12], to the best of our knowledge, no past work has studied using an external resource to improve the relevance estimate of document passages for ad hoc retrieval.

Our main contributions are: (1) We develop a new approach for representing passage-level evidence for ad hoc retrieval via a novel

[*] This author is also affiliated with Universitas Indonesia. This work was conducted during her graduate studies at RMIT University
[†] The research work was primarily done at RMIT University

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210028

use of CQA data; (2) The new approach provides a strong indication of document relevance, and is able to outperform many previous passage-based methods; combining text quality with evidence derived from the new representation leads to further improvements. Our experiments show that incorporating the new evidence significantly improves over state-of-the-art ranking models, including Quality-Biased Ranking (QSDM), External Expansion (EE), and a combination of both.

The remainder of this paper is organized as follows. Section 2 presents related work, followed by the motivation of this work in Section 3. Section 4 details our framework of passage extraction using external resources and document re-ranking using quality features derived from the answer passages. Sections 5 and 6 describe the experiment and the results. Discussion and concluding remarks are given in Sections 7 and 8.

2 BACKGROUND

Table 1 shows our categorization of approaches to document ranking: considering the object being scored (i.e. document or passage/summary) and the location of information that is exploited (i.e. local or external).[1] Work listed in the top-left cell focuses on attempts to improve relevance estimation of a document using the local collection. The top-right cell lists work exploring the use of more focused text representations such as passages or summaries. The bottom-left cell lists work exploiting external resources for improving relevance estimation. A considerable amount of effort was invested in both directions, but the intersection, the bottom-right, has had less exploration. We now examine each cell in turn.

Document-Based Scoring Using Local Collection. Common retrieval models such as BM25 [38], language models [34], and DfR [2] are in this cell. Among these widely used models, Sequential Dependency Model (SDM) has consistently demonstrated strong effectiveness [29]. Lavrenko and Croft [23] implemented pseudo-relevance feedback (PRF) within a language modeling framework. The basic idea of PRF [39] is assuming the initially retrieved top-ranked documents are relevant and then extracting the most frequent terms from them to improve retrieval effectiveness.

Kurland and Lee [20] leveraged link-based methods using inter-document similarities. Bendersky et al. [3] integrated document quality features in a quality-biased SDM (QSDM) framework, and showed the effectiveness of their approach over text- and link-based techniques. To the best of our knowledge, no previous work has reported superior performance to QSDM ranking.

Document-Based Scoring Using External Resources. Use of external resources to improve the relevance estimation of documents has also been tried [5, 12]. Diaz and Metzler [12] incorporated information from external corpora, such as the web collections, using a language modeling technique for PRF. External expansion was shown to be effective and was extended by Weerkamp et al. to the task of blog retrieval [45]. Bendersky et al. [5] also explored the use of term/concept statistics derived from external corpora, such as MSN query logs and Wikipedia, into the SDM method.

Passage-Based Scoring Using Local Collection. Combining evidence from passages to improve ad hoc retrieval has been explored by Bendersky and Kurland [4], who showed that incorporating the best-matching passage into the original document language model [34] can significantly improve retrieval effectiveness. Krikon and Kurland [19] further explored the integration of document-based, cluster-based, and passage-based information to improve document ranking. Relatively little work has considered using document summaries, as another passage representation, to improve retrieval effectiveness. More recently, He et al. [14] showed that combining summaries and documents improves the retrieval effectiveness of a document language model baseline.

However, as will be shown in our experiments, this approach does not improve over stronger retrieval models such as SDM, suggesting that the passage scoring is not effective in the presence of term proximity information. While the advantage of using passage representations in ad hoc retrieval appears evident, it is still an open question if further improvements can be made.

Passage-Based Scoring Using External Resources. To the best of our knowledge, using an external collection to better estimate the relevance of retrieved passages for ad hoc retrieval has not been explored. Much of the past work has focused on using external collections for the relevance estimation of documents, including Wikipedia [5, 28] and web collections [12].

CQA sites allow people to ask questions that are answered by other users in the community. The popularity of CQA, such as Yahoo! Answers, has grown rapidly; in 2016, over 3.0 million people in U.S. accessed Yahoo! Answers per month.[2] Previous work has exploited CQA data for many purposes, such as: answering factoid and tips questions [6, 44], answering non-factoid queries [51], predicting information asker and web searcher satisfaction [25, 26], and evaluating answer quality [41]. We are not aware of previous work that has used CQA for improving document ranking.

3 HYPOTHESIS

Our work tests the following *answer-bearingness* hypothesis:

> *Documents that are likely to bear focused answers to the posed query should be ranked highly.*

To test the hypothesis, CQA resources are exploited as proxies of an oracle "answer source", which is unattainable otherwise. A scoring rule is developed and used in a subsequent passage generation step to score any given passage according to how well its text content approximates the answer source data. Following Bendersky and Kurland [4], we assume that the best-scoring passage under this scoring rule can represent the full document in a quality-biased ranking framework, and therefore quality features derived from the best-scoring passage can directly benefit retrieval. A set of similar strategies was recently reviewed in passage retrieval [17], with an aim of improving the presentation of search results in general.

We now formally define the research questions:

RQ1 Can answer passages be exploited to improve document ranking compared to existing methods?

RQ2 Can incorporating quality features from answer passages improve ad hoc retrieval?

[1] Note, such a categorization excludes methods that either address more specific retrieval problems (e.g. clustering [36] or learning to rank [27]) or that exploit other data (e.g. link analysis [20] or user signals [1]).

[2] https://www.quantcast.com/answers.yahoo.com

Table 1: Ad hoc retrieval methodologies broken down in two axes, based on the object being scored (columns) and the resource used in relevance estimation (rows). Shaded methods are our addition to this work.

	Document	Passage
Local Collection	Retrieval models: BM25, SDM, or DfR Pseudo relevance feedback [23, 39] Quality-biased ranking (QSDM) [3]	Passage-based LM [4, 14, 19]
External Resources	External expansion (MoRM) [12, 45] Weighted dependence model (WSD) [5]	Answer-passage quality

Our methodology allows for the creation of multiple passage representations for improving document ranking, which leads to a third research question:

RQ3 Does combining quality features from multiple passage representations make a stronger ranking model?

4 AN ANSWER-PASSAGE APPROACH

In passage retrieval, a two-phase approach is used to avoid needing to generate passage representations for all documents. We assume that an initial set of documents $\mathcal{D}_Q$ with respect to query Q is first retrieved using a standard retrieval function such as BM25 or SDM, to serve as input to the passage retrieval module. Following this step, our answer-passage approach will exploit information from CQA data to induce passages that are likely to bear answers to query Q, and use this passage representation to re-rank documents.

We present two different methodologies in the coming sections for extracting and scoring answer passages. Section 4.1 presents a general probabilistic framework that involves external resources in the process of extracting answer passages. An alternative method, described in Section 4.2, leverages open-domain question answering to directly retrieve answer-reporting passages. On either type of representation, a final re-ranking step is performed based on the passage quality, which is described in Section 4.3.

4.1 A Probabilistic Framework

Our approach requires one basic functionality from the CQA resource: the ability to perform question retrieval so that the user can submit a query Q to retrieve a set of related questions and gain access to the respective answers $\mathcal{A}_Q$ (see Section 5.1). The answers, $\mathcal{A}_Q$, are used to improve the estimation of term relevance [12]. In a standard language modeling framework [23], this relevance estimate $p(t|Q)$ is written as:

$$p(t|Q) \propto \sum_{A\in\mathcal{A}_Q} p(t|A)\, p(Q|A), \tag{1}$$

where $p(t|A)$ is the relevance estimate of term t derived from answer A, and $p(Q|A)$ is the retrieval score of answer A with respect to Q.

Improving Relevance Estimation of Terms. For term relevance $p(t|A)$, we consider estimates that are in proportion to a given term weighting function. The following functions are discussed:

- Query Likelihood (QL) [34, 53]:

$$\frac{f(t,A) + \mu\, p(t|C)}{|A| + \mu}. \tag{2}$$

- BM25 [38]:

$$\frac{f(t,A)\,(k_1+1)}{f(t,A) + k_1\left(1 - b + b\,\frac{|A|}{\mathrm{avg}_{A'}|A'|}\right)}\, idf(t). \tag{3}$$

- Embedding Language Model (EMB):

$$\frac{\left[\prod_{t_A\in A} p(t,t_A)\right]^{1/|A|}}{\sum_{t'\in\tilde{\mathcal{T}}}\left[\prod_{t_A\in A} p(t',t_A)\right]^{1/|A|}}. \tag{4}$$

The first two functions are based on commonly used retrieval models, Query Likelihood (QL) [34, 53] and BM25 [38]. For QL, μ controls the degree of Dirichlet smoothing and $p(t|C)$ is the background (collection) language model. For BM25, k_1 and b are parameters and $\mathrm{avg}_{A'}|A'|$ is the average answer size. In these equations, $f(t,A)$ denotes the frequency of term t within answer A.

The third term relevance estimate is based on word embeddings [31], which can serve as an alternative to more conventional score functions. Our formulation differs from prior work [21, 52] in the way the probability of jointly observing term t and answer A is defined:

$$p(t,A) \propto \left[\prod_{t_A\in A} p(t,t_A)\right]^{1/|A|}. \tag{5}$$

We postulate that the likelihood of jointly observing two terms t and t' in the same document context is proportional to a sigmoidal transformation (with scale/location parameters κ and x_0) of the cosine similarity between the respective word vectors v_t and $v_{t'}$:

$$p(t,t') \propto \frac{1}{1 + \exp\left(-\kappa(\cos(v_t, v_{t'}) - x_0)\right)}. \tag{6}$$

It can be shown that the relevance estimate $p(t|A)$ as in (4) follows this derivation. Practically, it suffices to compute the normalization factor in (4) over a smaller subset of terms $\tilde{\mathcal{T}} \subset \mathcal{T}$.

$p(Q|A)$ informs the degree of relevance of answer A with respect to query Q, so that in (1) more relevant answers have stronger influence over the inferred model $p(t|Q)$. As CQA sites do not usually reveal such scores or even the scoring rules, some distributional assumptions are made for computing this estimate. One can assume that the query likelihood of answer A retrieved at the k-th position (within the set $\mathcal{A}_Q$) is distributed logarithmically, in accordance with the Discounted Cumulative Gain (DCG) [16], or geometrically, in accordance with the Rank-Biased Precision (RBP) metric [32]. For simplicity, in this paper we focus on only the DCG variant, defined as follows, as both variants showed comparable performance in our preliminary experiments:

$$p(Q|A) \propto (\log k + 1)^{-1}. \tag{7}$$

Extracting Answer Passages. The next step is to incorporate the estimated term relevance into a passage algorithm to extract sub-document representations G that best approximate the retrieved answer-bearing content. Two approaches are taken: extracting fixed-length passages (PSG) and extracting summaries using integer linear programming (ILP). Note that, depending on the approach in use, G can either be a contiguous block of text or a set of sentences put together by using document summarization.

The first approach, PSG, is based on the use of fixed-length passages that are common in retrieval [8, 33]. Such representations do not (usually) stick to predefined sentence/paragraph boundaries and can be easily generated using a sliding window algorithm. From all passages in document D, prior work [4] suggests scoring them with a language modeling approach to choose one passage G^* with the maximum score. Our first approach follows this practice but uses the improved relevance estimates to evaluate passages:

$$G^*_{\text{PSG}} = \arg\max_{G \in D} \sum\nolimits_{t \in G} p(t|Q). \tag{8}$$

However, the answer-bearing content may not necessarily form a contiguous text block so that fixed-length passages will catch them. Redundant terms in a passage can also fill up the space easily without providing additional information, rendering the relevance estimate unreliable.

Our second approach, ILP, draws on document summarization to tackle these issues. It leverages integer linear programming to extract document summaries [13, 42, 47], with the core algorithm extended to incorporate term relevance estimates derived from CQA resources. This particular approach is taken in our framework for both the efficacy and the ease to incorporate external knowledge about topical relevance.[3] The algorithm is optimized to select a set of sentences that maximize the coverage of answer-bearing terms in the generated summary G:

$$G^*_{\text{ILP}} = \arg\max_{G \in D} L(G) + \lambda R(G), \tag{9}$$

where:

$$L(G) = \sum_{t \in G} p(t|Q), \qquad R(G) = \sum_{t \in G} p(t|Q)\, sf(t, G) \tag{10}$$

and $|G^*|$ is less than or equal to some predefined K and $sf(t, G)$ denotes the "sentence frequency" of term t in summary G. Both objective functions $L(G)$ and $R(G)$ are combined using a hyperparameter $0 \le \lambda \le 1$. The first objective will try to maximize summary-level term coverage and reduce term repetition. The other sentence-level objective will include more sentences with highly relevant terms.

4.2 Open-Domain Question Answering

We also implement an alternative answer-passage scoring framework based on a recent open-domain question answering model, called Document Reader (DR) [9]. The goal of open-domain question answering is to automatically extract text fragments ("answers") from a set of unstructured or free-format documents to address users' questions.

Table 2: List of passage quality features.

Feature	Definition
`PassageScore`	Objective value to score the passage
`PassageOverlap`	Bigram overlap with respect to answers
`NumSentences`	Number of sentences
`QueryOverlap`	Number of query term occurrences
`AvgWordWeight`	Average passage term weight
`AvgTermLen`	Average passage term length
`Entropy`	Shannon entropy of the term distribution
`FracStops`	Fraction of passage terms that are stopwords
`StopCover`	Fraction of stopwords appear in the passage

The DR model takes query Q and document D as input and returns a best-matching passage $G^* = \langle g_1, g_2, \ldots, g_m \rangle$ of m terms that maximizes an *answer span* score, defined as follows:

$$G^*_{\text{DR}} = \arg\max_{G \in D} \max_{1 \le i \le j \le m} \log p_S(g_i|G, Q) + \log p_E(g_j|G, Q). \tag{11}$$

In this formulation, the score being optimized indicates the log-likelihood of a passage G reporting an answer. The core idea behind `DR` is to use recurrent neural networks to aggregate term-level evidence (i.e. features), and then for each passage term g_i estimate if the term starts or ends an answer span with respect to Q using attentive modeling [15]. The best scoring pair $\langle g_i, g_j \rangle$ in a passage is identified to compute the final answer span score. Specifically, the two likelihood models p_S and p_E, for starting and ending an answer span, are defined as:

$$\begin{aligned} p_S(g_i|G, Q) &\propto \exp(v_{g_i}^T\, W_S\, v_Q), \\ p_E(g_j|G, Q) &\propto \exp(v_{g_j}^T\, W_E\, v_Q), \end{aligned} \tag{12}$$

where v_{g_i} and v_{g_j} are passage-term vectors, v_Q denotes the query vector, and W_S and W_E indicate the bilinear mappings. Both passage-term vectors and query-term vectors are derived from the hidden states of two separate recurrent neural networks, and the query vector is a weighted combination of the derived query-term vectors. These definitions are given as follows:

$$\begin{aligned} \langle v_g : g \in G \rangle &= \text{BiLSTM}_G(\langle f_g : g \in G \rangle), \\ \langle v_q : q \in Q \rangle &= \text{BiLSTM}_Q(\langle e_q : q \in Q \rangle), \\ v_Q &= \sum_{q \in Q} \text{softmax}(W_Q\, v_q)\, v_q, \end{aligned} \tag{13}$$

where W_Q is a linear mapping, f_g denotes the feature vector for passage term g, and e_q denotes the word embeddings for term q.

4.3 Passage Quality Based Ranking

A mix of novel and existing features are employed to estimate the quality of the produced passage, see Table 2. `PassageScore` denotes the score assigned to the best matching passage in the retrieved document. The score is combined with `PassageOverlap` to estimate the answer-bearingness level of a passage relative to a given query. `PassageOverlap` measures the term overlap between a document passage and its related CQA answers. `NumSentences` is employed as a quality feature based on the idea that a summary with too many short sentences is less likely to be relevant or informative. `QueryOverlap` has been used in previous studies on web

[3] More advanced approaches, such as submodular optimization [24], use sentence-to-sentence similarities rather than concept relevance to perform document summarization. It is not clear yet how CQA resources can be incorporated in this regard to improve the extraction of answer passages.

Table 3: Test collections used in our experiments.

Collection	Topics	# Docs
GOV2	TREC Topics 701–850	25,205,179
ClueWeb09B	TREC Web Topics 1–200	50,066,642

search ranking [1] and in query-biased summarization [30]. Other prior work [50] leveraged `AvgWordWeight` as a sentence feature to generate document summaries. Motivated by the effectiveness of document quality features used by Bendersky et al. [3], we adopt four non-HTML specific quality features to work at the passage level: `AvgTermLen`, `Entropy`, `FracStops`, and `StopCover`.

The proposed quality estimates are combined by using a feature-based linear ranking model (see (14)). Previous work has used a similar approach [3, 29], and in most cases combining evidence from different representations and different retrieval functions has been shown to be beneficial [11]. As was done in the QSDM framework [3], the SDM retrieval score is also included in the model:

$$\lambda_D f_{\mathrm{SDM}}(q, D) + \sum\nolimits_j \lambda_j \, f_j(q, G) \qquad (14)$$

where the weights $\lambda_D + \sum_j \lambda_j = 1$, f_j represents the j-th feature, and G represents the answer passage. The weights are learned using a learning-to-rank algorithm described in Section 5.3.

5 EXPERIMENTS

A series of experiments was conducted to evaluate the effectiveness of the proposed ranking model using quality features extracted from the answer passages. Section 5.1 describes the data and evaluation metrics used in our experiments. Section 5.2 covers the details about baselines and Section 5.3 covers the parameter estimation.

5.1 Setup

The code and data used in this paper are made publicly available for interested readers to reproduce our results.[4]

Test Collections. Ranking experiments were conducted on two web test collections, GOV2 and CW09B (i.e. ClueWeb09B), using TREC Terabyte 2004–2006 and Web Track 2009–2012 “title” topics respectively. An overview of these data sets is provided in Table 3. Both web collections were indexed using the Indri search engine using Krovetz stemming without removing stopwords. The spam filter by Cormack et al. [10] was applied to CW09B, removing spam webpages with a score less than 70. Repeating the same experiments on un-filtered CW09B data leads to the same conclusions, with some slight decreases in absolute early precision (@10) but increases in recall-oriented metrics.

Retrieval Settings. Initially, a ranked list of 100 documents was retrieved using the SDM, following the configuration parameters suggested in the original paper $(\lambda_T, \lambda_O, \lambda_U) = (0.85, 0.10, 0.05)$ [29]. This step is performed using the Indri search engine.[5] The raw HTML content for each retrieved document was parsed by using BeautifulSoup[6] and sentences extracted using the Stanford CoreNLP toolkit.[7] Stopwords were removed from the sentences (using the INQUERY list) and Krovetz stemming was performed.

External CQA Resources. The external CQA data were obtained from Yahoo! Answers (Y!A), by submitting our queries to the Y!A search engine and taking the best answer for each of the top ten matching questions. In Y!A, the best answer for each question is chosen by the person who posts the question.[8] Our decision to use only the best answer for each question is to ensure good quality information [51]. There are three GOV2 queries (QID 703, 769, 816) and five CW09B queries (QID 95, 100, 138, 143, 146), however, that do not have any matching questions. Since the purpose of this research is to investigate how external evidence can be used to enhance summaries and document ranking, we remove these eight queries from this experiment (we return to the issue of the availability of suitable CQA answers in Section 6.5). The average number of related CQA answers per query in GOV2 and CW09B data are 9.52 and 9.74 (maximum of 10), respectively. The choice of ten as the number of related CQA answers per query is justified based on the result of an initial experiment, where we tried using 1, 5, 10, 20, 50, and 100, and found that according to several metrics, using a single answer is the least effective, while using ten answers gave the most effective results in most of the cases.

Word Embeddings. Two sets of word embeddings are used, both based on the fasttext package [7]. The first is a pre-trained set of one million word vectors based on the English Wikipedia data in 2017 of 16 billion tokens (denoted as EmbWiki), and the second is a set of five million word vectors trained on our custom crawl of Yahoo! Answers data of five billion tokens (denoted as EmbYA) using the skip-gram algorithm.[9] Both sets of vectors are of 300 and 100 dimensions respectively.

Evaluation Metrics. To get a broader understanding to the effectiveness of the proposed method, six evaluation metrics are reported in this study. Top-k effectiveness as the focus of web search is represented by NDCG@10, NDCG@20, P@10, and P@20. The metric MRR, which is widely used in web question answering, is also included. Additionally, we report MAP@100, as our ranking experiment is limited to the top 100 initially retrieved documents. The two-tailed t-test is used for significance testing.

5.2 Baselines

The following baselines were selected and implemented:

- Sequential Dependence Model (`SDM`);
- Passage-Based Language Model (`MSP` and `SUM`);
- Quality-Biased Ranking (`QSDM`);
- External Expansion (`EE`).

Passage-Based Language Models. A passage-based language model is a mixture of three models of the passage p_G, the document p_D, and the collection p_C. The combined model usually takes the following form:

$$p(Q) = \prod\nolimits_{t \in Q} \left[\lambda_G \, p_G(t) + \lambda_D \, p_D(t) + \lambda_C \, p_C(t)\right],$$

[4] https://github.com/rmit-ir/AnswerPassageQuality
[5] http://www.lemurproject.org/indri.php (version 5.9)
[6] https://www.crummy.com/software/BeautifulSoup/ (version 4.0)
[7] https://stanfordnlp.github.io/CoreNLP/ (version 3.8.0)
[8] http://yahooanswers.tumblr.com/post/80173794953/important-changes-to-answers
[9] The custom Yahoo! Answers crawl contains roughly 17 million question-answer pairs submitted to Yahoo! Answers between 2013 and 2016.

under the constraint that the mixture weights sum to one. The passage model p_G and the mixture weights might be implemented slightly differently across methods.

Two variants, MSP and SUM, are implemented in this paper. The first model is based on a top-performing variant MSP[length] from Bendersky and Kurland [4]. It locates the best-matching passage G in the document by maximizing the maximum-likelihood estimate p_G across a set of candidates, with one key parameter λ_D set by using the document homogeneity estimate $h^{[length]}$. A second approach, called SUM, based on query-biased summarization [14] was shown to be competitive to gradient boosting regression trees. Following the proposed setting [14], the MEAD package [35] is used to implement this method, combining four features: Centroid, Position, Length, and QueryCosine with the default weights.

Quality-Biased Ranking (QSDM). The quality-biased ranking method [3] is commonly referred to as the state of the art in web document ranking with TREC collections. The method is a linear model that combines the SDM score and ten web document quality features, which are: NumVisTerms, NumTitleTerms, AvgTermLen, FracAnchorText, FracVisText, Entropy (Entropy of the document content), FracStops, StopCover, UrlDepth (depth of the URL path), and FracTableText.

External Expansion (EE). External Expansion [12] is a standard PRF approach for expanding queries using external corpora based on the Relevance Model [23]. It is generally considered as a strong and effective expansion method when external resources are available.

5.3 Parameter Estimation

Parameters for individual baseline methods are tuned as follows:

- For passage-based language models (MSP and SUM), the mixture weights are optimized via a grid search over the range {0.00, 0.05, 0.10, ..., 0.95, 1.00} using cross validation.
- For external expansion (EE) the procedure followed closely to the original paper Diaz and Metzler [12]. The number of feedback documents (i.e., CQA answers) was set to ten to align with the data. The number of feedback terms n_T, collection model weight λ_C, and the mixture ratio λ_Q with respect to the original query were all learned on the target test collections via 100 rounds of randomized search over randomly re-sampled train/test (50%–50%) query splits.[10] In our experiments, $(n_T, \lambda_C, \lambda_Q)$ were set to $(60, 0.3, 0.2)$ on GOV2 and to $(50, 0.2, 0.2)$ on CW09B.

Parameters for experimental runs are tuned as follows:

- The passage size K is set to fifty words, to be made consistent with the common setting for query-biased summarization [35]. We set $\lambda = 0.1$ in the extraction of the ILP representation.
- For QL, we set $\mu = 100$ and for BM25, we set $b_1 = 1.2$ and $k_1 = 0.75$, based on common settings in adhoc retrieval. Both $p(t|C)$ and $idf(t)$ are estimated on the target collection.
- For both embedding based estimates EmbWiki and EmbYA, we set $\kappa = 10$ and $x_0 = 0$ based on cross validation.
- Our implementation of the DR framework follows the original paper [9]: we encode query and passage vectors using 128 hidden units in three-layer bidirectional LSTMs. The model is trained on the SQuAD dataset [37] using AdaMax [18]. The dropout rate is tuned empirically to 0.5. We use the same set of word embeddings learned from the Y!A data (as with EmbYA), but the effectiveness is roughly comparable to a pre-trained model learned on the Common Crawl data [9].

For all methods tested in our experiments, a Coordinate Ascent learning-to-rank algorithm is employed to learn the model weights using ten-fold cross validation, as is commonly practiced in past work [3, 29].[11] We used RankLib[12] to estimate parameters, which are essentially the weight of each feature. We chose to optimize NDCG@20 throughout the experiments as it gives the best performance in terms of both precision- and recall-oriented metrics.

6 RESULTS

We describe and analyze the effectiveness of ranking using different answer-passage representations: PSG and ILP, as well as passages derived by using open-domain question answering model (DR).

6.1 Comparisons with Previous Work

Our approach is first compared with prior techniques for both test collections, see Table 4. Ten experiments are reported: two representations PSG and ILP with four relevance estimates EmbWiki, EmbYA, QL, and BM25, and the DR framework are tested using title and description queries.

It can be seen that combining SDM with answer-passage quality using all three representations PSG, ILP, and DR, significantly outperforms SDM and the passage-based baselines SDM+MSP and SDM+SUM. While incorporating MSP and SUM shows only marginal benefits over SDM, combining answer-passage quality has seen the biggest effect across all the other methods involved. This provides an answer to RQ1: *answer passages can be used to improve ad hoc retrieval in the presence of a strong retrieval baseline* SDM*, and they can work better than existing passage-based methods.*

The fact that DR does not provide strong indication of document relevance is unexpected. Among all representations ILP is found to be the most effective, while PSG and DR are roughly comparable. For both SDM+PSG and SDM+ILP, BM25 gives the best results and QL the second, followed by embedding based estimates EmbWiki and EmbYA. The SDM+DR framework works the best on description queries, suggesting that further tuning might be needed for such models to handle non-verbose queries. On both test collections, the best effectiveness is achieved by using SDM+ILP with BM25.

6.2 Ad Hoc Retrieval with Passage Quality

The previous experiment shows that SDM+ILP paired with retrieval function BM25 and QL may have some advantages over strong retrieval model QSDM and SDM+EE. This leads to a further investigation

[10] This optimization procedure allows the resulting retrieval scores to be included as a feature in our ranking model. The resulting scores are comparable to the procedure proposed by Diaz and Metzler [12].

[11] Note that the comparison between ranking algorithms is beyond the scope of this paper. In preliminary experiments, non-linear ranking models such as Gradient Boosted Decision Trees (GBDT) and LambdaMART were also tested, but were found to consistently perform less effectively than Coordinate Ascent on all metrics by a wide margin, suggesting that the ideal response surface is close to a hyperplane, as non-linear models can struggle with this type of ranking problem.

[12] https://www.lemurproject.org/ranklib.php (version 2.7)

Table 4: Comparisons with previous methods. Significant differences with respect to SDM/QSDM/SDM+EE are indicated using †/⋄/∗ for p < 0.05 (or ‡/⋄⋄/∗∗ for p < 0.01). All differences between SDM+PSG/ILP runs and SDM+MSP are significant for p < 0.05.

		NDCG@10	NDCG@20	P@10	P@20	MRR	MAP@100
	Baseline / Passage Baseline						
	SDM$^{(\dagger)}$	0.4769	0.4751	0.5694	0.5469	0.7763	0.1802
	QSDM$^{(\diamond)}$	0.5127‡	0.5022‡	0.6197‡	0.5759‡	0.8174†	0.1919‡
	SDM+EE$^{(*)}$	0.5189‡	0.5057‡	0.6129‡	0.5738‡	0.8220†	0.1879‡
	SDM+MSP	0.4826	0.4745	0.5782	0.5422	0.7696	0.1805
	SDM+SUM	0.4741	0.4749	0.5680	0.5500	0.7729	0.1805
	Answer-Passage Approach						
GOV2	SDM+PSG (EmbWiki)	0.4999‡	0.4975‡	0.6041‡	0.5745‡	0.8063	0.1888‡
	SDM+PSG (EmbYA)	0.5010†	0.4957†	0.6007†	0.5724‡	0.8024	0.1888‡
	SDM+PSG (QL)	0.5085‡	0.5068‡	0.6102‡	0.5823‡	0.7991	0.1929‡
	SDM+PSG (BM25)	0.5174‡	0.5116‡	0.6184‡	0.5847‡	0.8271‡	0.1946$^{\ddagger *}$
	SDM+ILP (EmbWiki)	0.5081‡	0.4967‡	0.6204‡	0.5752‡	0.8098	0.1892‡
	SDM+ILP (EmbYA)	0.4983†	0.4951‡	0.6075‡	0.5779‡	0.7900	0.1876‡
	SDM+ILP (QL)	0.5131‡	0.5052‡	0.6238‡	0.5844‡	0.7878	0.1964$^{\ddagger **}$
	SDM+ILP (BM25)	**0.5293**‡	**0.5171**‡	**0.6367**‡	**0.5946**$^{\ddagger *}$	**0.8234***	**0.2009**$^{\ddagger \diamond **}$
	SDM+DR (Title)	0.4821	0.4786	0.5735	0.5480	0.7817	0.1811
	SDM+DR (Desc)	0.4999‡	0.4894†	0.6014‡	0.5612	0.8038	0.1842$^{\dagger \diamond\diamond}$
	Baseline / Passage Baseline						
	SDM$^{(\dagger)}$	0.2542	0.2462	0.3682	0.3321	0.5010	0.1053
	QSDM$^{(\diamond)}$	0.2735	0.2639†	0.3938†	0.3467	0.5224	0.1094
	SDM+EE$^{(*)}$	0.2880‡	0.2736‡	0.4021‡	0.3590‡	0.5619‡	0.1136‡
	SDM+MSP	0.2535	0.2469	0.3656	0.3328	0.4989	0.1054
	SDM+SUM	0.2499	0.2409	0.3631	0.3267	0.4952	0.1047
	Answer-Passage Approach						
CW09B	SDM+PSG (EmbWiki)	0.2693†	0.2588†	0.3831	0.3421	0.5325	0.1058
	SDM+PSG (EmbYA)	0.2752†	0.2644†	0.3856	0.3479	0.5299	0.1103
	SDM+PSG (QL)	0.2613	0.2569	0.3805	0.3490†	0.5222	0.1087
	SDM+PSG (BM25)	0.2811†	0.2687‡	0.3938†	0.3521†	0.5499†	0.1113†
	SDM+ILP (EmbWiki)	0.2843†	0.2652†	0.3954	0.3392	0.5803‡	0.1070
	SDM+ILP (EmbYA)	0.2818‡	0.2665†	0.3990‡	0.3485	0.5579†	0.1092
	SDM+ILP (QL)	0.3090$^{\ddagger \diamond\diamond}$	0.2901$^{\ddagger \diamond\diamond}$	0.4313$^{\ddagger \diamond\diamond *}$	0.3736$^{\ddagger \diamond\diamond}$	0.5786$^{\ddagger \diamond}$	0.1164$^{\ddagger \diamond}$
	SDM+ILP (BM25)	**0.3115**$^{\ddagger \diamond\diamond *}$	**0.2955**$^{\ddagger \diamond\diamond *}$	**0.4379**$^{\ddagger \diamond\diamond *}$	**0.3787**$^{\ddagger \diamond\diamond}$	**0.5902**$^{\ddagger \diamond\diamond}$	**0.1209**$^{\ddagger \diamond\diamond *}$
	SDM+DR (Title)	0.2584	0.2505	0.3662	0.3295	0.5298	0.1050
	SDM+DR (Desc)	0.2833‡	0.2681‡	0.3949†	0.3441	0.5613‡	0.1094

regarding improvements over strong retrieval models. We next incorporate passage quality features into an expanded set of retrieval models, using the ILP representation together with BM25 and EmbYA relevance estimates. For the choice of base models, we used SDM, QSDM, and QSDM+EE, with the latter being a novel and strong combination of quality-biased ranking and external expansion.

The results (Table 5) show three rows in each collection for each base model. Incorporating ILP significantly improves SDM for all metrics, across collections. On GOV2, BM25 improves over QSDM for NDCG@10, NDCG@20 and MAP@100. On CW09B, using BM25 leads to significant increases over QSDM for all metrics. For QSDM+EE, significant increases were observed on P@10, P@20, and MAP@100 on the GOV2 data using BM25, and CW09B runs show a similar trend but with a more pronounced effect. We conclude that RQ2 is answered: *incorporating answer-passage quality can significantly improve ad hoc retrieval in general, but as the base system improves, further gains are likely to get smaller.*

6.3 Combining Multiple Representations

Next, two answer-passage representations are involved in the ranking process. Denoted as Combined, this new experimental run effectively leverages passage-level evidence from two representations learned by using different methodologies. The aim is to understand whether quality estimates derived from different representations provide similar effects to document ranking.

For this experiment, we incorporate answer-passage quality estimates from both representations ILP (BM25) and DR (Desc) into the

Table 5: Retrieval effectiveness of ranking models using quality estimates of answer-biased summaries. Significant differences with respect to baselines SDM/QSDM/QSDM+EE are indicated using †/⋄/∗ for $p < 0.05$ (or ‡/⋄⋄/∗∗ for $p < 0.01$).

		NDCG@10	NDCG@20	P@10	P@20	MRR	MAP@100
GOV2	SDM$^{(\dagger)}$	0.4769	0.4751	0.5694	0.5469	0.7763	0.1802
	SDM+ILP (EmbYA)	0.4983†**	0.4951‡**	0.6075‡*	0.5779‡	0.7900**	0.1876$^{\ddagger\diamond**}$
	SDM+ILP (BM25)	0.5293‡	0.5171‡	0.6367‡	0.5946‡	0.8234†	**0.2009**$^{\ddagger\diamond*}$
	QSDM$^{(\diamond)}$	0.5127‡	0.5022‡	0.6197‡	0.5759‡	0.8174†	0.1919‡
	QSDM+ILP (EmbYA)	0.5197‡	0.5126‡	0.6238‡	0.5874‡	0.8258†	0.1891‡*
	QSDM+ILP (BM25)	0.5412$^{\ddagger\diamond}$	0.5245$^{\ddagger\diamond\diamond}$	0.6463‡	0.5939‡	0.8338†	0.2007$^{\ddagger\diamond\diamond*}$
	QSDM+EE$^{(*)}$	0.5339$^{\ddagger\diamond}$	0.5213$^{\ddagger\diamond\diamond}$	0.6374‡	0.5901‡	**0.8416**‡	0.1948‡
	QSDM+EE+ILP (EmbYA)	0.5329$^{\ddagger\diamond}$	0.5208$^{\ddagger\diamond}$	0.6429‡	0.5959$^{\ddagger\diamond}$	0.8044*	0.1947‡
	QSDM+EE+ILP (BM25)	**0.5442**$^{\ddagger\diamond\diamond}$	**0.5311**$^{\ddagger\diamond\diamond}$	**0.6605**$^{\ddagger\diamond\diamond*}$	**0.6082**$^{\ddagger\diamond\diamond*}$	0.8407‡	0.1996$^{\ddagger\diamond\diamond**}$
CW09B	SDM$^{(\dagger)}$	0.2542	0.2462	0.3682	0.3321	0.5010	0.1053
	SDM+ILP (EmbYA)	0.2818‡*	0.2665†*	0.3990‡	0.3485	0.5579†	0.1092*
	SDM+ILP (BM25)	0.3115$^{\ddagger\diamond\diamond}$	0.2955$^{\ddagger\diamond\diamond}$	0.4379$^{\ddagger\diamond\diamond*}$	0.3787$^{\ddagger\diamond\diamond}$	0.5902$^{\ddagger\diamond\diamond}$	0.1209$^{\ddagger\diamond\diamond*}$
	QSDM$^{(\diamond)}$	0.2735	0.2639†	0.3938†	0.3467	0.5224	0.1094
	QSDM+ILP (EmbYA)	0.2853†	0.2691†	0.3923	0.3485	0.5566†	0.1109
	QSDM+ILP (BM25)	0.3107$^{\ddagger\diamond\diamond}$	0.2959$^{\ddagger\diamond\diamond}$	0.4333$^{\ddagger\diamond\diamond*}$	0.3774$^{\ddagger\diamond\diamond}$	0.6002$^{\ddagger\diamond\diamond}$	0.1190$^{\ddagger\diamond\diamond}$
	QSDM+EE$^{(*)}$	0.2985$^{\ddagger\diamond}$	0.2819‡+	0.4056‡	0.3610‡	0.5799$^{\ddagger\diamond\diamond}$	0.1148$^{\ddagger\diamond}$
	QSDM+EE+ILP (EmbYA)	0.3042$^{\ddagger\diamond\diamond}$	0.2864$^{\ddagger\diamond\diamond}$	0.4174‡	0.3679$^{\ddagger\diamond}$	0.5881$^{\ddagger\diamond\diamond}$	0.1169$^{\ddagger\diamond}$
	QSDM+EE+ILP (BM25)	**0.3194**$^{\ddagger\diamond\diamond*}$	**0.3015**$^{\ddagger\diamond\diamond**}$	**0.4338**$^{\ddagger\diamond**}$	**0.3826**$^{\ddagger\diamond\diamond**}$	**0.6138**$^{\ddagger\diamond\diamond}$	**0.1210**$^{\ddagger\diamond\diamond**}$

Table 6: Combining ILP and DR significantly improves QSDM (significant differences are indicated using ⋄ for $p < 0.05$ or ⋄⋄ for $p < 0.01$).

		N@20	P@20	MAP@100
GOV2	QSDM$^{(\diamond)}$	0.5022	0.5759	0.1919
	QSDM+Combined	0.5280$^{\diamond\diamond}$	0.6007$^{\diamond}$	0.1972
CW09B	QSDM$^{(\diamond)}$	0.2639	0.3467	0.1094
	QSDM+Combined	0.2896$^{\diamond\diamond}$	0.3656$^{\diamond}$	0.1166$^{\diamond}$

QSDM run. The results of these experiments are shown in Table 6. The Combined method produces strong retrieval runs, but not significantly better than just incorporating ILP. However, QSDM+Combined significantly outperforms QSDM on both collections, and across all metrics except MRR and MAP@100 on GOV2. Regarding the answer to RQ3, we conclude that *there is some evidence to support the claim that the use of multiple representations will lead to a stronger ranking model.*

6.4 Feature Importance

An ablation analysis was conducted on the run QSDM+EE+ILP (BM25) with twenty one features in total, to examine the relative feature importance. The top seven features for each collection are shown in Table 8, ordered by decreasing difference of NDCG@20 score after removing a feature. NDCG@20 is used as an ordering criterion following our optimization metric in the main experiment. The letter P in square brackets indicates passage-level quality features.

SDM remains the most important feature across collections. Some differences between collections can be seen based on the relative importance of features. Our passage quality features AvgWordWeight, QueryOverlap, and FracStop[P] appear more effective on GOV2. On CW09B, PassageScore, StopCover[P], and AvgTermLen are among the top-ranked features. Note that PassageScore and EE also show high importance on CW09B, indicating the usefulness of external CQA resources in ad hoc retrieval.

6.5 Lack of CQA Resources

To investigate to what extent the ranking effectiveness changes when the coverage of good related CQA answers is not guaranteed, we conduct an experiment using the related answers obtained from the offline Yahoo! Webscope L6 collection[13] using a mixture approach [49].[14] Note that this dataset was collected prior to the creation of CW09B, so the respective TREC query topics are less likely to have direct answers in the data.

Our results (Table 9) suggest a decrease of up to 2.6% on GOV2 and 3% on CW09B compared to the result of using the related CQA answers obtained from Yahoo! Answers (Y!A) search engine (see Table 5). The implication is that a good coverage of related answers is crucial in the generation of answer passages, that are primarily supported by the size of collection and the human efforts involved in curating the best answers.

It is worth noting that using an offline resource with limited coverage of related CQA answers still leads to a significant improvement in NDCG@20 and P@20 on the CW09B collection. This

[13] http://webscope.sandbox.yahoo.com/

[14] The weight of SDM retrieval score for question title, question body, and best answer fields are respectively set to 0.5, 0.2, and 0.3 based [49].

Table 7: Answer passages extracted from a top-ranked relevant document clueweb09-enwp01-16-17964 for TREC Web Topic 65, "*Find information and resources on the Korean language.*" (query: korean language)

ILP (EmbYA)	For example, different endings are used based on whether the subjects and listeners are friends, parents, or honoured persons. in a similar way European languages borrow from Latin and Greek. Its use limited some cases and the aristocracy prefers Classical Chinese for its writing. "Mortal enemy" and "head of state" are homophones in the South. Learn to read, write and pronounce Korean
ILP (BM25)	Yanbian (People's Republic of China) Given this, it is sometimes hard to tell which actual phonemes are present in a certain word. Unlike most of the European languages, Korean does not conjugate verbs using agreement with the subject, and nouns have no gender. The Korean language used in the North and the South exhibits differences in pronunciation, spelling, grammar and vocabulary.
DR (Desc)	Korean is similar to Altaic languages in that they both lack certain grammatical elements, including number, gender, articles, fusional morphology, voice, and relative pronouns (Kim Namkil). Korean especially bears some morphological resemblance to some languages of the Northern Turkic group, namely Sakha (Yakut).

Table 8: Results of ablation study to determine feature importance for both test collections.

GOV2		*CW09B*	
Feature	Diff.	Feature	Diff.
SDM	0.0306	SDM	0.0223
FracStop	0.0101	StopCover	0.0092
AvgWordWeight	0.0076	PassageScore	0.0086
UrlDepth	0.0063	FracVisText	0.0077
QueryOverlap	0.0052	EE	0.0076
FracAnchorText	0.0049	StopCover[P]	0.0046
FracStop[P]	0.0048	AvgTermLen	0.0038

Table 9: An investigation of using external CQA resources from offline collection. Significant differences with respect to QSDM are indicated using $\diamond$ for $p < 0.05$ (or $\diamond\diamond$ for $p < 0.01$).

		N@20	P@20	MAP@100
GOV2	QSDM$^{(\diamond)}$	0.5022	0.5759	0.1919
	QSDM+ILP (BM25)	0.5083	0.5759	0.1926
CW09B	QSDM$^{(\diamond)}$	0.2639	0.3467	0.1094
	QSDM+ILP (BM25)	0.2804$^{\diamond\diamond}$	0.3679$^{\diamond\diamond}$	0.1136

is in line with one previous study [51] on exploiting CQA resources for non-factoid question answering, which shows modest improvements in the quality of produced answers even when no CQA answer exactly matches the queries. We note that, in the case where CQA answers are not available for a particular query, a live system can simply back off to a retrieval mode that does not incorporate such evidence, as the lack of appropriate CQA resources is clearly indicated through an empty results list.

7 DISCUSSION

The results in the previous sections provide strong empirical evidence to support the validity of the answer-bearingness hypothesis, and also directly support the recurring argument in previous work [4, 8, 14, 22, 40, 46] that passage-level evidence can benefit retrieval effectiveness. It is however surprising that, the open-domain question answering model shows little benefit in extracting answer-bearing passages for document ranking. This may be due to task mismatch (i.e. model trained to detect factoids) or the lack of appropriate training instances. Word embeddings learned on the CQA data are arguably useful for the task, but simpler methodologies appear to win on the overall efficacy.

The ILP representation benefits the most from the inclusion of passage quality estimates, which we suspect is due to the fact that summaries are more likely to cover broken sentences on non-relevant documents. The compressive nature forces summaries to include all textual evidence that seems relevant, but when such evidence is scarce the quality can be poor.

For illustrative purposes, some example answer passages produced by using ILP and DR are also given in Table 7. The answer passages are extracted from a top-ranked relevant document for a randomly sampled query topic. We note that the extraction algorithms tend to capture a broader range of answers when the underlying document is relevant, as is shown in the example. While the captured evidence may differ in terms of efficacy for document ranking, it remains an open question how this difference correlates with users' perception towards the answer-passage quality.

8 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed quality-biased ranking that incorporates signals from passages that are likely to bear answers. A new approach that exploits external resources in the creation of such passages is developed to induce high-quality sub-document representations, called answer passages, from the retrieved documents. We developed a set of methodologies to improve term relevance estimates and extract answer passages. A range of quality features is extracted from the generated passages, and blended into the ranking model, which leads to improved effectiveness: our experiments on two web collections showed that this approach is more effective than passage-based methods and external expansion, and can significantly improve on state-of-the-art ranking models SDM and QSDM. Signals from multiple representations can also be combined to improve ranking effectiveness. A final ranking model that combines all these quality estimates achieved significant effectiveness improvements on GOV2 and ClueWeb09B.

In future work we plan to conduct a user study to gain an understanding of human perceptions of the quality of answer passages, and of the improvement in document ranking. A promising new approach to entity representation has recently been published [48], which approaches the problem of ranking from an angle orthogonal to our work. We plan to explore a combination of such approaches in future. We will look to examine possible improvements to our ranking model, such as using link-based features [20], user behavior signals [1], and filtering CQA answers based on their quality [41].

9 ACKNOWLEDGMENTS

This research is supported in part by the Australian Research Council (DP140102655 and DP170102726) and the Indonesia Endowment Fund for Education (LPDP). The authors want to thank Marwan Torki for kindly sharing his CQA research data.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proc. of SIGIR*. ACM, 19–26.

[2] Gianni Amati and Cornelis Joost van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.* 20, 4 (2002), 357–389.

[3] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. 2011. Quality-biased Ranking of Web Documents. In *Proc. of WSDM*. ACM, 95–104.

[4] Michael Bendersky and Oren Kurland. 2008. Utilizing passage-based language models for document retrieval. In *Proc. of ECIR*. Springer, 162–174.

[5] Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2010. Learning Concept Importance Using a Weighted Dependence Model. In *Proc. of WSDM*. ACM, 31–40.

[6] Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *Proc. of WWW*. ACM, 467–476.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[8] James P. Callan. 1994. Passage-level Evidence in Document Retrieval. In *Proc. of SIGIR*. Springer-Verlag New York, Inc., 302–310.

[9] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proc. of ACL*. Association for Computational Linguistics, 1870–1879.

[10] Gordon V. Cormack, Mark D. Smucker, and Charles L. Clarke. 2011. Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets. *Inf. Retr.* 14, 5 (Oct. 2011), 441–465.

[11] W Bruce Croft. 2002. Combining approaches to information retrieval. In *Proc. of ECIR*. Springer, 1–36.

[12] Fernando Diaz and Donald Metzler. 2006. Improving the Estimation of Relevance Models Using Large External Corpora. In *Proc. of SIGIR*. ACM, 154–161.

[13] Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*. Association for Computational Linguistics, 10–18.

[14] Jing He, Pablo Duboue, and Jian-Yun Nie. 2012. Bridging the Gap between Intrinsic and Perceived Relevance in Snippet Generation. In *Proc. of COLING*. 1129–1146.

[15] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. 1693–1701.

[16] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[17] Mostafa Keikha, Jae Hyun Park, and W Bruce Croft. 2014. Evaluating answer passages using summarization measures. In *Proc. of SIGIR*. ACM, 963–966.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Eyal Krikon and Oren Kurland. 2011. A study of the integration of passage-, document-, and cluster-based information for re-ranking search results. *Information Retrieval* 14, 6 (2011), 593–616.

[20] Oren Kurland and Lillian Lee. 2010. PageRank without hyperlinks: Structural reranking using links induced by language models. *ACM TOIS* 28, 4 (2010), 18.

[21] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query expansion using word embeddings. In *Proc. of CIKM*. ACM, 1929–1932.

[22] Adenike M. Lam-Adesina and Gareth J. F. Jones. 2001. Applying Summarization Techniques for Term Selection in Relevance Feedback. In *Proc. of SIGIR*. ACM, 1–9.

[23] Victor Lavrenko and W Bruce Croft. 2001. Relevance based language models. In *Proc. of SIGIR*. ACM, 120–127.

[24] Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Proc. of HLT/NAACL*. Association for Computational Linguistics, 912–920.

[25] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. 2011. Predicting web searcher satisfaction with existing community-based answers. In *Proc. of SIGIR*. ACM, 415–424.

[26] Yandong Liu, Jiang Bian, and Eugene Agichtein. 2008. Predicting information seeker satisfaction in community question answering. In *Proc. of SIGIR*. ACM, 483–490.

[27] Craig Macdonald, Rodrygo L.T. Santos, and Iadh Ounis. 2012. On the Usefulness of Query Features for Learning to Rank. In *Proc. of CIKM*. ACM, 2559–2562.

[28] Edgar Meij and Maarten de Rijke. 2010. Supervised query modeling using wikipedia. In *Proc. of SIGIR*. ACM, 875–876.

[29] Donald Metzler and W. Bruce Croft. 2005. A Markov Random Field Model for Term Dependencies. In *Proc. of SIGIR*. ACM, 472–479.

[30] Donald Metzler and Tapas Kanungo. 2008. Machine Learned Sentence Selection Strategies for Query-Biased Summarization. In *SIGIR Learning to Rank Workshop*.

[31] Bhaskar Mitra and Nick Craswell. 2017. Neural Models for Information Retrieval. *arXiv preprint arXiv:1705.01509* (2017).

[32] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.* 27, 1 (2008), 2.

[33] John O'Connor. 1980. Answer-passage retrieval by text searching. *Journal of the Association for Information Science and Technology* 31, 4 (1980), 227–239.

[34] Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. of SIGIR*. ACM, 275–281.

[35] Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. MEAD — A platform for multidocument multilingual text summarization. In *Proc. of LREC*.

[36] Fiana Raiber and Oren Kurland. 2013. Ranking document clusters using markov random fields. In *Proc. of SIGIR*. ACM, 333–342.

[37] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[38] Stephen E Robertson. 1997. Overview of the okapi projects. *Journal of Documentation* 53, 1 (1997), 3–7.

[39] Joseph John Rocchio. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing* (1971), 313–323.

[40] Tetsuya Sakai and Karen Sparck-Jones. 2001. Generic Summaries for Indexing in Information Retrieval. In *Proc. of SIGIR*. ACM, 190–198.

[41] Chirag Shah and Jefferey Pomerantz. 2010. Evaluating and predicting answer quality in community QA. In *Proc. of SIGIR*. ACM, 411–418.

[42] Hiroya Takamura and Manabu Okumura. 2009. Text summarization model based on maximum coverage problem and its variant. In *Proc. of EACL*. Association for Computational Linguistics, 781–789.

[43] Anastasios Tombros and Mark Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proc. of SIGIR*. ACM, 2–10.

[44] Ingmar Weber, Antti Ukkonen, and Aris Gionis. 2012. Answers, not links: extracting tips from yahoo! answers to address how-to web queries. In *Proc. of WSDM*. ACM, 613–622.

[45] Wouter Weerkamp, Krisztian Balog, and Maarten de Rijke. 2012. Exploiting External Collections for Query Expansion. *ACM Trans. Web* 6, 4 (2012), 1–29.

[46] Ross Wilkinson. 1994. Effective Retrieval of Structured Documents. In *Proc. of SIGIR*. Springer-Verlag New York, Inc., 311–317.

[47] Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proc. of EMNLP*. Association for Computational Linguistics, 233–243.

[48] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. 2017. Word-Entity Duet Representations for Document Ranking. In *Proc. of SIGIR*. ACM, 763–772.

[49] Xiaobing Xue, Jiwoon Jeon, and W Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proc. of SIGIR*. ACM, 475–482.

[50] Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In *Proc. of SIGIR*. ACM, 255–264.

[51] Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2018. Document summarization for answering non-factoid queries. *IEEE Trans. Knowl. Data Eng.* 30, 1 (2018), 15–28.

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

[53] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (2004), 179–214.