Passage Level Evidence for Effective Document Level Retrieval

Ghulam Sarwar[1], Colm O'Riordan[1] and John Newell[2]
[1]*Department of Information Technology, National University of Ireland, Galway, Ireland*
[2]*School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland*

Keywords: Document Retrieval, Passage-based Document Retrieval, Passage Similarity Functions.

Abstract: Several researchers have considered the use of passages within documents as useful units of representation as individual passages may capture accurately the topic of discourse in a document. In this work, each document is indexed as a series of unique passages. We explore and analyse a number of similarity measures which take into account the similarity at passage level with the aim of improving the quality of the answer set. We define a number of such passage level approaches and compare their performance. Mean average precision (MAP) and precision at k documents (P@k) are used as measures of the quality of the approaches. The results show that for the different test collections, the rank of a passage is a useful measure, and when used separately or in conjunction with the document score can give better results as compared to other passage or document level similarity approaches.

1 INTRODUCTION

Information Retrieval (IR) deals with the organization, representation, and the retrieval of information from a large set of text documents. The retrieval of relevant information from large collections is a difficult problem; search queries and documents are typically expressed in natural language which introduces many problems such as ambiguity caused by the presence of synonyms and abbreviations, and issues arising from the *vocabulary difference problem* which occurs when the user expresses their information need with terms different to those used to express the same concept in the document collection.

Several models have been shown to be very effective in ranking documents in terms of their relevance to a user's query. The user formulates the query by expressing their information need in natural language. Approaches include different mathematical frameworks (vector space model, probabilistic models) to represent documents and queries and to formulate a comparison approach. The BM25 weighting scheme (Robertson et al., 2009) derived within a probabilistic framework is a well-known effective one in estimating the relevance of a document to a query. The main goal of an IR system is to estimate the relevance of a document to a query; this notion of 'relevance' is often interpreted as measuring the level of similarity between a document to a query.

In IR, the traditional approaches consider the document as a single entity. However, some researchers choose to split the document into a separate passages given the intuition that a highly relevant passage may exist in a larger document which itself will be considered as non relevant. If a passage is indexed as an individual *pseudo-document*, the number of documents stored and indexed will increase significantly and in a result, it will effect the speed and cost of retrieval (Roberts and Gaizauskas, 2004). However, one may now retrieve relevant passages that occur in documents deemed not very relevant. Moreover, if the document returned as relevant is too long, it can be difficult for the users to find the appropriate relevant passages in the document. In other words, returning a large relevant document, while useful, still, puts an onus on the user to find the relevant passages. Therefore, we opt for the passage level retrieval approach to finding the relevant passage and aim to use that to improve the document ranking. The intuition behind our approach is that by identifying very relevant passages in a document we can better estimate the relevance of the overall document.

One can imagine the passages themselves as documents at indexing time. The division of these

Sarwar G., O'Riordan C. and Newell J.
Passage Level Evidence for Effective Document Level Retrieval.
DOI: 10.5220/0006502800830090
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*, pages 83-90
ISBN: 978-989-758-271-4
Copyright © 2017 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

passages can be done in a number of ways. For example, either via some textual identifier e.g. paragraph markings ($< p >$), new line feed (/n) etc. or it can be defined by a number of words. A passage could be a sentence, a number of sentences or a paragraph itself. The passages can be considered as discrete passages with no intersection or can be viewed as overlapping passages.

In this paper we introduce different similarity functions that were used to generate new document rankings by computing the passage similarity and using this score (or its combination with document level similarity score) as a means to rank the overall document.

The main focus of our work is to see how effectively the passage level evidence affected the document retrieval. Factors such as different means to define passage boundaries are not of huge concern to us as present.

We have used the WebAp (Web Answer Passage)[1] test collection which is obtained from the 2004 TREC Terabyte Track Gov2 collection and the Ohsumed test collection (Hersh et al., 1994) which comprises titles and/or abstracts from 270 Medline reference medical journals. The results show that different similarity functions behave differently across the two test collections.

The paper outline is as follows: section 2 presents a brief overview of the previous work in passage level retrieval. Section 3 gives an overview of the methodology employed, outlining the details of different similarity functions, the passage boundary approach, and the evaluation measures adopted in the experiments. Section 4 presents a brief explanation of the test collections used in the experiments and the assumptions made for them. Section 5 discusses different experimental results obtained. Finally, section 6 provides a summary of the main conclusions and outlines future work.

2 RELATED WORK

In previous research, passage level retrieval has been studied in information retrieval from different perspectives. For defining the passage boundaries, several approaches have been used. Bounded passages, overlapping window size, text-tiling, usage of language models and arbitrary passages (Callan, 1994; Hearst, 1997; Bendersky and Kurland, 2008b; Kaszkiel and Zobel, 2001; Clarke et al., 2008) are among the few main techniques. Window size approaches consider the word count to separate the passages from each other, irrespective of the written structure of the document. Overlapping window size is shown to be more effective and useful for the document retrieval (Callan, 1994). Similarly, a variant of the same approach was used by Croft (Liu and Croft, 2002).

Jong (Jong et al., 2015) proposed an approach which involved considering the score of passages generated from an evaluation function to effectively retrieve documents in a Question Answering system. Their evaluation function calculates the proximity of the different terms used in the query with different passages and takes the maximum proximity score for the document ranking.

Callan (Callan, 1994) demonstrated that ordering documents based on the score of the best passage may be up to 20% more effective than standard document ranking. Similarly, for certain test collections, it was concluded that combining the document score with the best passage score gives improved results. Buckley et al also use the combination of both scores in a more complex manner, to generate scores for ranking (Buckley et al., 1995). Moreover, Hearst et al (Hearst and Plaunt, 1993) showed that instead of only using the best passage with the maximum score, adding other passages gives better overall ranking as compare to the ad-hoc document ranking approach.

Salton (Salton et al., 1993) discussed another idea to calculate the similarity of the passage to the query. They re-ranked and filtered out the documents that has a low passage score associated with it. They included all the passages that have a higher score than its overall document score, and then used these scores to raise, or lower, the final document rank. In this way, the document that has a lower score to the document level score but a higher score at passage level for certain passages, will get a better ranking score in the end.

Different language modelling approaches at passage level and document level have been used in the past to improve the document ranking (Liu and Croft, 2002; Lavrenko and Croft, 2001). A similar approach has been used by Bendersky et al (Bendersky and Kurland, 2008b), where they used the measure of the document homogeneity and heterogeneity to combine the document and passage similarity with the query

[1] https://ciir.cs.umass.edu/downloads/WebAP/

to retrieve the best documents. To use the passage level evidence, their scoring method used the maximum query-similarity score that is assigned to any passage in the document ranking. As for their passage based language model, they used the simple unigram based standard to estimate the probabilities at passage and document level. Moreover, Krikon and Kurland (Krikon et al., 2010; Bendersky and Kurland, 2008a) used a different language modeling approach where they tried to improve the initial ranking of the documents by considering the centrality of the documents and the passages by building their respective graphs. The edges denote the inter-term similarities and the centrality is computed using the page rank approach. They reported that their approach performed better than the normal maximum passage approach and some variation of interpolation score of maximum passage score with document score.

3 METHODOLOGY

In traditional adhoc IR, a 'bag of words' model is adopted with no attention paid to word order or word position within a document. Weights are typically assigned to terms according to some heuristics, probability calculations or language model.

In this work, we view every document as being represented as passages or 'pseudo-documents' i.e. $d^{'} = \{p_1, p_2, \ldots p_n\}$. We attempt to better estimate $sim(d,q)$ by estimating $sim(d^{'},q)$. Different similarity functions are designed in a way that different characteristics of the passage level results can be used alone, or in combination with the document level results. We define $sim(d^{'},q)$ as $f(sim(p_i,q), sim(d,q))$

3.1 Similarity Functions

Following is a brief description of these similarity functions in which different characteristics were computed from the passage level evidence:

- {**SF1**} Max Passage: One way to compute the $sim(d^{'},q)$ is to consider the similarity and ranking of the passage that has the highest similarity score to the query as a representative of the similarity of the document.

$$sim(d^{'},q) = max(sim(p_i,q))$$

- {**SF2**} Sum of passages: It is similar to the max passage approach, but instead of taking only the top passage, the top k of the passages are taken and their similarity scores are combined by adding them together.

$$sim(d^{'},q) = \sum_{i=1}^{k}[sim(p_i,q)]$$

- {**SF3**} Combination of document and passage similarity scores: In this case, the passage and document scores are combined and then the results are re-ranked based on the new score.

$$sim(d^{'},q) = \alpha(max(sim(p_i,q))) + \beta(sim(d,q))$$

- {**SF4**} Inverse of rank: Rather than using the document or passage scores, the rank at which these passages are returned can also be used to find the similarity between the passages and the query. This can be calculated as follows:

$$sim(d^{'},q) = (\frac{\sum_i \frac{1}{rankP_i}}{\#of\, p_i}) \quad |p_i \in d^{'}$$

- {**SF5**} Weighted Inverse of Rank: Another way to take the rank of these passages into account is to take the sum of the inverse ranks and pay less attention to lower ranks. Hence, the higher ranks will impact more on the results as compare to the lower values and will effect the overall ranking.

$$sim(d^{'},q) = \sum_i (\frac{1}{rankP_i})^{\alpha} \quad |p_i \in d^{'}, \alpha > 1$$

3.2 Passage Boundaries

To run the experiments, all the documents and passages were first indexed in our IR system. We have used Solr 5.2.1[2] as a baseline system which is a high performance search server built using Apache Lucene Core. In this system, a vector space model is adopted with a weighting scheme based on the variation of tf-idf and Boolean model (BM) (Lashkari et al., 2009) is used.

We use two different test collections in the experiments. The WebAP test collection contains 6399 document and 150 queries in its dataset. We adopt overlapping windows for this collection and decompose each document into passages of length 250 words. This results in the creation of 140,000 passages for the WebAP collection. The second collection, the Ohsumed dataset, comprises 348,566 Medline abstracts as documents with 106 search queries. Given the relatively small document lengths, in defining passage boundaries, an overlapping window size of 30 words is used for this collection which creates a document set of passages of size 1.4 million pseudo-documents that gives 4-5 passages per document. We choose the half overlapping, fixed length window-size to index the documents, because

[2] http://lucene.apache.org/solr/5_2_1/index.html

these passages are more suitable computationally, convenient to use, and were proved to be very effective for document retrieval(Callan, 1994; Liu and Croft, 2002).

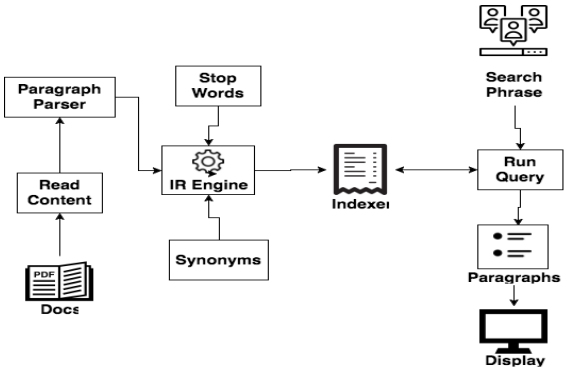


Figure 1: Architectural Diagram.

3.3 Evaluation

To evaluate the results and measure the quality of our approach, mean average precision (MAP) and precision@k are used as the evaluation metrics. The MAP value is used to give an overall view of the performance of the system with different similarity functions. Furthermore, precision@k was helpful in illustrating the behavior of the system with respect to correctly ranking relevant documents in the first k positions.

4 EXPERIMENTAL SETUP

In this section, we present a brief explanation of the test collections we used, and also some detail of different parameters that we consider in our experiments. Lastly we will describe the brief overview of the evaluation measures that we used in the experiments.

4.1 Test Collections

For our experiments we used the two different test collections that are freely available to use for experimental purposes.The following is a brief explanation of both datasets.

4.1.1 WebAp

Web Answer Passage (WebAP) is a test collection, which is obtained from the 2004 TREC Terabyte Track Gov2 collection. The dataset contains 6399 documents and 150 query topics and relevance judgment of top 50 documents per query topic. It is created mainly for the purpose of evaluating passage level retrieval results (Keikha et al., 2014) but has been used in question answering (QA) task to retrieve sentence level answers as well (Chen et al., 2015; Yang et al., 2016). The query topic section contains keyword based queries and the normal queries. We generated the results against both types and here we reported the performances that are based on the keyword based queries. On average, these results performed overall 2% better than the normal query ones across all similarity functions. Annotation at passage level (GOOD, FAIR, PERFECT etc.) is also included in this test collection that can be used to differentiate the different passages in term of their relevance to the query. The annotators found 8027 relevant answer passages to 82 TREC queries, which is 97 passages per query on average. From these annotated passages, 43% of them are perfect answers, 44% are excellent, 10% are good and the rest are fair answers. We have saved these passage annotations while indexing them in the system, but, we have not used them in our evaluation criteria. As the size of all the documents are fairly large compare to the other test collections we came across, therefore, we divided passages using overlapping window based approach of size 250 words.

4.1.2 OHSUMED

The Ohsumed collection consists of titles and abstracts from 270 Medline reference medical journals. It contains 348,566 articles along with 106 search queries. In total, there are 16,140 query-documents pairs upon which the relevance judgments were made. These relevance judgments are divided in three categories i.e. definitely relevant, possibly relevant, or not relevant. For experiments and evaluation, all the documents that are judged here as either possibly or definitely relevant were considered as relevant. Furthermore, only the documents to which the abstracts are available, were index and used for the retrieval task. Therefore, the experiments were conducted on the remaining set of 233,445 documents from the Ohsumed test collection. Also, to calculate the overall performance we considered only those queries, which had any relevant document(s) listed in the judgment file. Out of 106 queries in total, 97 of them were found to have relevant document(s) associated with it. This document collection is fairly large in terms document size but shorter in terms of document length as compare to the WebAP test collection. It does not include any annotation at passage level.

Table 1: MAP(%) For WebAp and Ohsumed Collection at k=5 and k=10.

Similarity Functions	MAP@5(WebAP)	MAP@10(WebAP)	MAP@5(Ohsumed)	Map@10(Ohsumed)
Document Level(D)	9.52	18.60	2.97	4.75
Max Passage(SF1)	9.43	18.56	3.23	4.96
Sum of Passages(SF2)	9.42	18.54	3.19	**4.99**
Inverse of Rank(SF4)	9.42	18.56	**3.27**	4.89
Weighted Inverse of Rank(SF5)	9.43	18.58	3.20	4.98
D+SF1	9.53	18.65	3.01	4.90
D+SF2	9.53	**18.67**	2.82	4.74
D+SF4	9.54	18.66	2.88	4.80
D+SF5	**9.55**	**18.67**	2.80	4.60

4.2 Assumptions and Experimental parameters

For our experiments we used Solr-5.2.1 which is built on top of LUCENE[3]. Solr provided the functionality of removing the stop-words at indexing time. As shown in figure 1, we used that functionality to remove the stop words[4] from both collections. We have seen that the ranking after removing the stop-words is improved.

For different similarity measure functions, we used different parameters. For sum of passages(SF2) and inverse rank(SF4) function, we set the k value to be equal to 5 and the results were normalized having received the final score. Similarly, we gave twice the boost to the passage level score as compared to the document level score while combining the results together i.e $\alpha = 1, \beta = 2$. Giving the higher boost to passage level gives better performance to the inverse ranking functions, whereas higher boost at document level improved results for Max passage and Sum of passage results.

4.3 Evaluation Measures

In IR, different evaluation measures are used to measure how well the system is performing to satisfy the user's need in returning the relevant documents to a given query. In our case, to measure the quality and performance of our approach, we used Mean Average Precision (MAP) and precision@k. MAP value is used to give an overall performance overview of the system and different similarity functions across both test collections. On the other hand, precision@k was helpful in illustrating the user's experience and the behavior of relevant documents returned in terms of their ranking frequency with the different threshold values. We evaluated the precision value for top 40 unique documents, both at passage level and at document level.

[3] http://lucene.apache.org/

[4] http://www.ranks.nl/stopwords

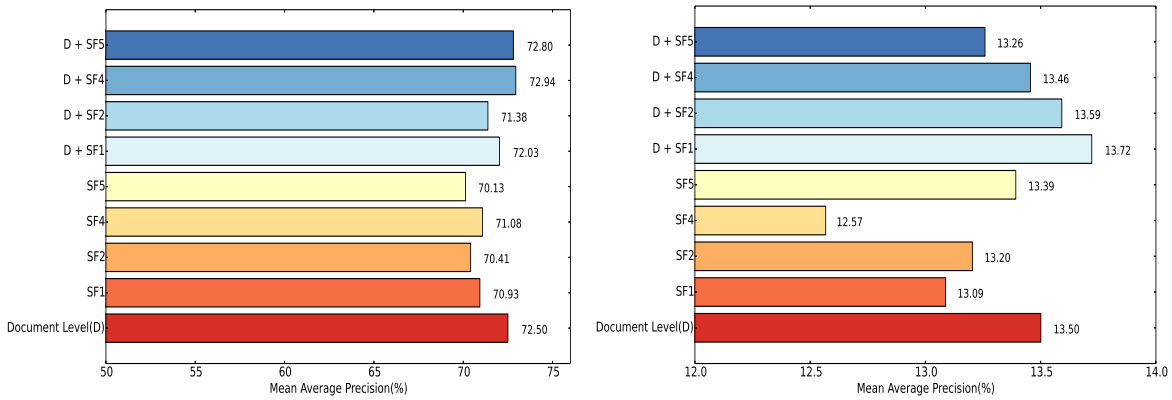
5 RESULTS

In this section we present the experimental results to show the performance of the different similarity functions at passage level and document level for both the WebAP and Ohsumed datasets.

In Figure 2(a) and 2(b), a bar chart is used to compare the document-level score with the different similarity functions of passage level scores for WebAP and Ohsumed test collections.

Using the WebAP collection, the results show that combining the document level score with passage level score (SF3), gives an improvement in performance. The best results were found when the document level score was combined with the inverse rank functions (SF4, SF5) of the passage level ranking. The results show that, considering the rank of the documents instead of the similarity score gives better performance when document ranking is combined with the passage level evidence. For the sum of passages (SF2) approach, only the top 5 (i.e. k=5) results were considered in calculating the query similarity score.

In contrast to WebAP, for the Oushmed collection the combination of document score with the max passage score performed better than the combination of inverse passage rank with document score. However, for functions not including the document level similarity, inverse rank by alpha (SF5) performed better than the other passage level similarity functions and give approximately similar performance in comparison to document level. Furthermore, the sum of passages (SF2) performed better here than the Max passage (SF1) score. The best results were observed for k=2. We have observed that the MAP values decrease as the k value increases, hence max passage similarity function performs better than the sum of passages function for WebAP test collection. However, in Ohsumed SF2 performed better than SF1 for $k = \{2,3,4\}$.



(a) WebAp Collection

(b) Ohsumed Collection

Figure 2: Mean Average Precision for Different Similarity Functions.

We also used precision@k as a different evaluation metric. The objective of this experiment was to check how well the documents are returned at the top k ranks at the document and passage level, and to measure on average how many relevant documents are returned at the different k values. Figure 3(a) and figure 3(b) illustrate the calculated precision values for WebAP test collection and the Ohsumed collection at document level as well as at passage level. At passage level we used SF4 and SF5 to measure the average precision for the WebAP and the Ohsumed, as when we considered it separately (without in conjunction with the document score), their performance was better than SF1 and SF2.

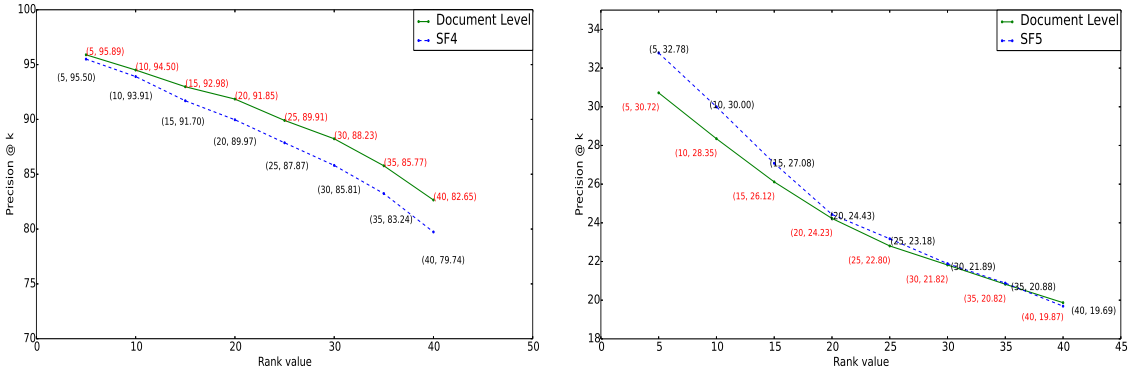
For the WebAP, the results show that the document level achieved better $p@k$ in comparison to SF4, and out of 40 documents, 33 of them are relevant in document level and 31 of them are relevant at the passage level when SF4 was used. On average, the precision value for document level and passage level was 90% and 86%. This indicates that the correct documents for all queries are clustered together or are closely related to each other and therefore, most of them are returned in top results, hence the high results.

For the Ohsumed collection, SF5 clearly outperformed the document level results and gave marginally better precision from the start to top 20 results (p@20) compared to the document level. However, for the higher values i.e. k>20 , the document level and SF5 gave almost the similar performance. Out of 40 documents approximately 9 are relevant in document retrieval and 10 of them are relevant in passage retrieval by using the inverse rank by alpha function(SF5). The overall performance for the Ohsumed collection is fairly low and this could be partially due to the large size of the test collection, small document length and the variation of relevant document information in relevance judgment file. On average, precision value for the document level and passage level was 24% and 25%.

Table 1 illustrates the mean average precision at top 5 (MAP@5) and at top 10 (MAP@10) for both test collections and as the results were discussed before, in the WebAP the combination of document level with passage level scores with different similarity functions give better results. The best results were obtained when the document score is combined with SF5. Whereas, for the Ohsumed, the functions that do not involve combining passage level and document level evidence gives better performance in both cases.

To get a better understanding on the statistical significance of the differences shown in the Table 1 for the test collections, we used the Student's t-test on paired samples for the top 50 MAP values with the difference of 5 (i.e. top 5, top 10, top 15, till top 50 etc). For the WebAP, we compared the document level results with the D+SF5 similarity function as it gave an overall better performance on the top results. The average MAP difference between both experiments was 0.18 with the standard deviation of 0.09 and the calculated p-value was 0.00024. Therefore, the performance shown by D+SF5 is statistically significant as compared to the normal document level results. Similarly, we performed the same t-test on the Ohsumed collection by comparing the document level results with D+SF4 due to its advantage over the performance on normal document level results. For the Ohsumed, the average difference and standard deviation were 0.07 and 0.13 with the p-value of 0.069. Hence, for the Ohsumed, the results were not improved very significantly.

It is also seen that the value of α and β effects the overall results when the document level is combined with the passage level evidence (SF3). For both collections, giving the higher boost to passage level i.e. $\alpha <= \beta$, gave a better performance for the inverse



(a) WebAP Collection (b) Ohsumed Collection

Figure 3: Precision at K for Different Test Collections.

ranking functions, whereas a higher boost at document level i.e. $\alpha > \beta$ improves the results for SF1 and SF2. We chose $\alpha = 1$ and $\beta = 2$ for the results shown in this paper because it gives an overall better performance for all the passage level similarity functions when combined with the document score.

6 CONCLUSIONS AND FUTURE WORK

In this paper our aim was to attempt to improve the document ranking by exploring and analyzing different similarity measures which take into account the similarity at passage level for all the documents. We used two different test collections; WebAP and Ohsumed, to measure the impact of a number of similarity functions at document level. The results shows that the rank of a passage is an effective measure and produced better results as compared to the functions that consider only the document score (i.e SF1, SF2). Furthermore, when it is combined with the document score, the performance was marginally improved for the WebAp test collection. The combination of the document score with the max passage score gave the best results for the Ohsumed collection. However, the improvements are minimal.

As per our results, it is shown that the passage level evidence on its own is not sufficient to improve the document ranking significantly for the selected test collections. Therefore, in the future, we are aiming to measure the impact of similarity functions listed in this paper on a fairly large size test collections such as TIPSTER, GOV2 and AQUAINT, in order to observe a significant different, if any. Additionally, we are aiming to explore different learning measures that we can use to combine the most suitable parameters for our interpolation equation i.e. α, β, score, rank etc. and report any significant improvements. Similarly, we are hoping to augment the passages with the different ontologies and words (entities, keywords, concepts etc.) related to the main keywords used in the query and the returned passages by using Wordnet or other knowledge based tools to improve the performance.

ACKNOWLEDGEMENTS

This work is supported by the Irish Research Council Employment Based Programme.

REFERENCES

Bendersky, M. and Kurland, O. (2008a). Re-ranking search results using document-passage graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 853–854. ACM.

Bendersky, M. and Kurland, O. (2008b). Utilizing passage-based language models for document retrieval. In *European Conference on Information Retrieval*, pages 162–174. Springer.

Buckley, C., Salton, G., Allan, J., and Singhal, A. (1995). Automatic query expansion using smart: Trec 3. *NIST special publication sp*, pages 69–69.

Callan, J. P. (1994). Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310. Springer-Verlag New York, Inc.

Chen, R.-C., Spina, D., Croft, W. B., Sanderson, M., and Scholer, F. (2015). Harnessing semantics for answer sentence retrieval. In *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 21–27. ACM.

Clarke, C. L., Cormack, G. V., Lynam, T. R., and Terra, E. L. (2008). Question answering by passage selection. In *Advances in Open Domain Question Answering*, pages 259–283. Springer.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Hearst, M. A. and Plaunt, C. (1993). Subtopic structuring for full-length document access. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 59–68. ACM.

Hersh, W., Buckley, C., Leone, T., and Hickam, D. (1994). Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR94*, pages 192–201. Springer.

Jong, M.-H., Ri, C.-H., Choe, H.-C., and Hwang, C.-J. (2015). A method of passage-based document retrieval in question answering system. *arXiv preprint arXiv:1512.05437*.

Kaszkiel, M. and Zobel, J. (2001). Effective ranking with arbitrary passages. *Journal of the American Society for Information Science and Technology*, 52(4):344–364.

Keikha, M., Park, J. H., Croft, W. B., and Sanderson, M. (2014). Retrieving passages and finding answers. In *Proceedings of the 2014 Australasian Document Computing Symposium*, page 81. ACM.

Krikon, E., Kurland, O., and Bendersky, M. (2010). Utilizing inter-passage and inter-document similarities for reranking search results. *ACM Transactions on Information Systems (TOIS)*, 29(1):3.

Lashkari, A. H., Mahdavi, F., and Ghomi, V. (2009). A boolean model in information retrieval for search engines. In *Information Management and Engineering, 2009. ICIME'09. International Conference on*, pages 385–389. IEEE.

Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127. ACM.

Liu, X. and Croft, W. B. (2002). Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382. ACM.

Roberts, I. and Gaizauskas, R. (2004). Evaluating passage retrieval approaches for question answering. In *European Conference on Information Retrieval*, pages 72–84. Springer.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Salton, G., Allan, J., and Buckley, C. (1993). Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58. ACM.

Yang, L., Ai, Q., Spina, D., Chen, R.-C., Pang, L., Croft, W. B., Guo, J., and Scholer, F. (2016). Beyond factoid qa: effective methods for non-factoid answer sentence retrieval. In *European Conference on Information Retrieval*, pages 115–128. Springer.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS