Exploiting structure and content of Wikipedia for Query Expansion in the context of Question Answering

Surya Ganesh, Vasudeva Varma
Language Technologies Research Centre,
IIIT-Hyderabad, India
suryag@research.iiit.ac.in, vv@iiit.ac.in

Abstract

Retrieving answer containing passages is a challenging task in Question Answering. In this paper we describe a novel query expansion method which aims to rank the answer containing passages better. It uses content and structured information (link structure and category information) of Wikipedia to generate a set of terms semantically related to the question. As Boolean model allows a fine-grained control over query expansion, these semantically related terms are added to the original query to form an expanded Boolean query. We conducted experiments on TREC 2006 QA data. The experimental results show significant improvements of about 24.6%, 11.1% and 12.4% in precision at 1, MRR at 20 and TDRR scores respectively using our query expansion method.

1 Introduction

Question Answering (QA) aims at finding exact answers to natural language questions in a large collection of documents (such as World Wide Web). Within a QA system, passage retrieval reduces the search space for finding the answer from such large collection of documents to a fixed number of passages (say top 20). But this could lead to the case where the set of passages considered may not contain the answer. In such cases any QA system would not answer the question. One of the reasons for not retrieving answer containing passages is attributed to the problem of vocabulary mismatch i.e., passages holding the answer to a question have semantic alterations of original terms in the question. Moldovan [5] showed that their system failed to answer 25.7% of questions solely because of vocabulary mismatch. In Information Retrieval (IR) such a problem is addressed using the technique of Query Expansion. It is the process of expanding the search query to match additional relevant documents. Query expansion techniques like adding synonyms of the query terms and relevance feedback have performed well in many IR applications. But most of these techniques as described by Derczynski [4] were not successful in the context of QA.

In this paper we describe a novel query expansion method using Wikipedia. Wikipedia is the leading open encyclopedia with a wide coverage on diverse topics, events, entities, etc. It is a reliable data-source and has found its use in many applications [1]. Another factor which motivated us to use it, is based on the simple experiment we conducted using TREC 2006 QA test set [3]. The test set consists of question series where each series asks for information regarding a particular target. The targets in the test set include people, organizations, events and other entities. Because of low data redundancy in Wikipedia, the coverage of its articles is directly proportional to the size of the text content in them. So in this experiment, we search for size of the text content present in Wikipedia for each target and the results are shown in table 1.

Target	Count
Rich Content	64
Partial Content	8
Zero Content	3

Table 1: *Targets from TREC 2006 QA test set*

As most of the targets as seen in Table 1, have rich content in Wikipedia, we used it as a knowledge source for our Query expansion method. Apart from the content of Wikipedia, we also used its structured information in our method. The structured data we used includes category information and link structure. Each article in Wikipedia belongs to one or more categories and the links between articles signify a semantic relationship between source and target articles.

2 Related Work

Different query expansion methods have been studied to enhance the performance of passage retrieval in the context of QA. Monz [6] tested a blind relevance feedback technique which selects terms based on standard Rocchio term weighting from top N documents. His experiments in the context of QA, showed a reduction in the performance compared to original query's performance. On the other hand, the same technique was found effective for the ad-hoc retrieval task. Pizzato [7] employed a blind relevance feedback technique which uses the named entities of the relevant answer type from the top ranked documents to form an expanded query. His experiments on PERSON type factoid questions have not shown a considerable improvement.

Yang et al. [10] used WordNet and Web to expand queries for QA. Only marginal improvements were attained when Web was used to extract expansion terms and when WordNet was used to rank these extracted terms the improvement was reduced. On semantic grouping the candidate expansion terms based on the relations between them, best results were obtained. Bilotti et al. [2] studied the effect of stemming and explicit query expansion using inflectional variants on document retrieval in the context of QA. The experimental results showed high recall for explicit query expansion and comparably low recall when stemming was used. Sun et al. [8] studied two query expansion techniques which make use of dependency relation analysis to extract contextual terms and relations from external corpuses. These techniques were used to enhance the performance of density based and relation based passage retrieval frameworks. The experimental results showed that relation based term expansion method with density based passage retrieval system outperformed the local content analysis method for query expansion. And, relation expansion method outperformed relation based passage retrieval system.

Arguello [1] described a technique for mining the links and anchor text in Wikipedia for query expansion terms and phrases. The technique yielded consistent and significant improvements in both recall and precision for blog recommendation. Our query expansion method is intended to rank the answer containing passages higher, by using the content and structure of Wikipedia.

3 Methodology

Our query expansion method first defines a query expansion term space ($QETS$) and then selects terms in this space based on proximity between terms and category information of outlink pages in Wikipedia. The query expansion term space consists of terms which could enhance the performance of passage retrieval for a given question. So, defining $QETS$ plays a major role in query expansion methods and it depends on different factors. In the case of Document Retrieval, query expansion methods are intended to bridge the gap between a high level general topic (expressed by the query) and the more nuanced facets of that topic likely to be written about in the documents. So, they use terms from top ranked documents or user selected documents to form $QETS$ for a given query. But, in the case of QA, query expansion methods are intended to rank the answer containing passages higher, as only fixed number of top ranked passages are considered to find the answer. So, constructing $QETS$ with the terms that are semantically related to the question could help in ranking the answer containing passages higher.

We use the content of Wikipedia to define $QETS$ for a given question in the following way. First, the Wikipedia article (A) corresponding to the question target is found and then a set of sentences (S) from this article which consist of question keywords is found. The above process of retrieving relevant sentences to a question is similar to that of passage retrieval. The terms in these sentences excluding stopwords and question keywords, constitute to form $QETS$ for a given question. It also includes terms in the anchor text of outlinks from the relevant sentences. Each term in this $QETS$ is weighted based on its semantic relatedness to the question. And, the strength of this semantic relation is captured using a linear combination of proximity score and outlink score as shown in the equation below.

$$score(t \in QETS) = ps(t, Q) + ls(t, C) \qquad (1)$$

Where t is a term in $QETS$, Q is a string of keywords in the question, C is the category information of outlink page, $ps(t, Q)$ and $ls(t, C)$ are proximity and outlink scores of term t. The significance and computation of proximity and outlink scores are described below.

3.1 Proximity score

The assumption behind selection of terms based on proximity scores is that semantically related terms are usually located in proximity, and the distance between two terms could indicate the strength of their association. The proximity score of a term is computed using its frequency and its minimum distance to a keyword in the question over a fixed window size of single sentence. Normally, within a sentence most of the terms occur only once. So, effectively our proximity score of a term is the summation of its minimum distances to a keyword in the question over all the relevant sentences (S) found in Wikipedia. Finally, each term in $QETS$ is weighted using the equation below.

$$ps(t \in QETS, Q) = \sum_{i=1}^{|S|} tf_{s_i}(t) * \frac{1}{dt_{s_i}(t, Q)}$$

Where $tf_{s_i}(t)$ is the term frequency of t in the sentence s_i and $dt_{s_i}(t, Q)$ is the minimum distance between term t and a keyword from Q.

3.2 Outlink score

This scoring method exploits the structured information (link structure and category information) of Wikipedia to rank the terms in $QETS$. All the outlinks present in the relevant sentence set (S) may not be semantically related to the question. So to find only the semantically related outlinks to the question, the category information of these outlink pages is used. Only those outlinks with their category information matching the question are considered semantically relevant. For example, given the question "*Which position did Warren Moon play in professional football?*", only the outlinks that fall into any one of these categories "*position/play/football/professional*" are considered semantically relevant to the question. Finally, all the terms from anchor texts of these relevant outlinks are weighted based on their frequencies in the relevant sentences (S) as shown in the equation below. And, for the rest of the terms in $QETS$, the outlink score is zero.

$$ls(t \in QETS, C) = tf_S(t)$$

The final scores of all the terms in $QETS$ is computed using equation 1, and they are sorted based on these scores. After sorting, the top N terms are picked for query expansion. The top 10 query expansion terms for the sample question "*Which position did Warren Moon play in professional football?*" from TREC 2006 QA dataset are shown in Table 2. One of the query expansion terms "*quarterback*" is the name of a position in football and even all other terms are semantically related to the keywords in the question. We use the term expansion length (el) which defines the number of terms considered for query expansion, in the rest of this paper. To trade off the balance between length of original query and expansion length, the latter must be proportional to the number of terms in the former.

$$el = k * |Q| \qquad (2)$$

Where, k is a constant and $|Q|$ is number of terms in the query. So, for short queries the expansion length will be small and for long queries the expansion length will be large.

quarterback	surpassed
league	canadian
american	record
completions	attempts
touchdowns	unmatched

Table 2: *Top 10 expansion terms for the question "Which position did Warren Moon play in professional football?".*

Boolean model allows a fine-grained control over query expansion. Tellex [9], in his study of different passage retrieval algorithms found that Boolean querying schemes perform well in the QA task. So, we use Boolean model to form the expanded query from the original query with appropriate weights. The expanded Boolean query is a combination of question target, keywords in the question and expansion terms from Wikipedia. Finally, the expanded Boolean query is given to the passage retrieval which searches for relevant paragraphs that are likely to contain the answer.

4 Experiments

The three principal measures used to measure the performance of passage retrieval in the context of QA are: Precision at 1, Mean Reciprocal Rank (MRR) at N, and Total Document Reciprocal Rank (TDRR). Precision at 1 is the proportion of questions for which a correct answer appears in the first retrieved passage. The MRR at N is the mean of the inverse of highest ranked correct answer if that answer appears in the top N. TDRR extends MRR with a notion of recall. It is the sum of all reciprocal ranks of all answer bearing passages per question (averaged over all questions) and attains maximum if all retrieved passages are relevant. In the experiments described below we considered top 20 passages for evaluation i.e. both MRR at N and TDRR are measured for top 20 passages.

We tested our query expansion method on TREC 2006 QA test set. The test set consists of AQUAINT corpus: contains 1,033,461 documents taken from the New York Times, the Associated Press, and the Xinhua News Agency newswires; question set: contains a series of 75 targets and each target contains a minimum of five factoid questions; answer judgments: contains answer patterns and document IDs in which they occur. TREC also provides the top 1000 documents for every target in the question set. These documents are retrieved from the AQUAINT collection using Prise[1] search engine. In our experiments we use Wikipedia dump as of October 13, 2008. The dump consists of about 4.0 million articles in XML format. We use Lucene[2] (a freely available open-source IR engine) for indexing and searching the Wikipedia articles. Lucene supports a Boolean query language, although it performs ranked retrieval using BM25. So, we used Lucene for retrieving relevant passages from top 1000 documents set in our experiments.

We conducted three experiments to test our query expansion method and in each experiment we did two separate evaluations, with strict and lenient criteria. For a passage to be judged correct within the strict criteria, the answer pattern must occur in the passage, and the passage must be from a document listed as relevant in the answer judgments. Under the lenient criteria, the answer pattern must occur in the passage. Strict scoring suffers from false negatives i.e., valid answer containing passages are scored as incorrect, since the list of document IDs supplemented in answer judgments is not exhaustive, and lenient scoring suffers from false positives i.e., wrong answer containing passages are scored as correct, since some of the answer patterns are not discriminating enough. So, strict and lenient scoring measure lower and upper bound performance of passage retrieval.

In the first experiment, we compared the performance of passage retrieval using expanded queries with the expansion length of $k = 8$ (in equation 2), against the one which is using seed/original queries. The results for this experiment over all the factoid questions in the test set are shown in Table 3. These results show improvements of about 24.6%, 11.1% and 12.4% in precision at 1, MRR at 20 and TDRR scores respectively under strict criteria and improvements of about 18.4%, 10.5% and 13.8% under lenient criteria.

Criteria	Metric	SQ	EQ
Strict	Prec@1	0.158	0.197
	MRR@20	0.252	0.280
	TDRR	0.330	0.371
Lenient	Prec@1	0.282	0.334
	MRR@20	0.387	0.428
	TDRR	0.742	0.845

Table 3: *Strict and lenient evaluation results for seed queries (SQ) and expanded queries (EQ)*

In the second experiment, we analyzed the two scoring methods to show how much does each of the two scores contribute to the overall performance of passage retrieval. For each question in the test set, two expanded queries with the expansion length of $k = 8$ (in equation 2) were constructed, where expansion terms for first and second queries were selected from $QETS$

[1] http://www-nlpir.nist.gov/works/papers/zp2/psearch_design.html

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

by using proximity and outlink scores respectively. The performance of passage retrieval using the above two queries are shown in Table 4. Comparing these results with seed and expanded queries performance from Table 3 shows an increase in the performance over the former but not as much as latter. So, the linear combination of proximity and outlink score results in better ranking of terms in $QETS$. And, in between the two scoring methods, outlink scoring performs better under strict criteria and proximity scoring performs better under lenient criteria.

Criteria	Metric	PS	OS
Strict	Prec@1	0.174	0.192
	MRR@20	0.262	0.279
	TDRR	0.351	0.373
Lenient	Prec@1	0.313	0.298
	MRR@20	0.413	0.396
	TDRR	0.825	0.786

Table 4: *Statistical analysis of proximity scoring (PS) and outlink scoring (OS) methods*

Finally, we tested our methodology for different expansion lengths by varying k value in the equation 2. Figure 1 shows the performance of passage retrieval for different expansion lengths under strict and lenient criteria. For both the criteria, the performance of our methodology has improved for all expansion lengths corresponding to k (in equation 2) values from 1 to 10 over the baseline ($k = 0$), and it attains maximum for the expansion length with $k = 8$.

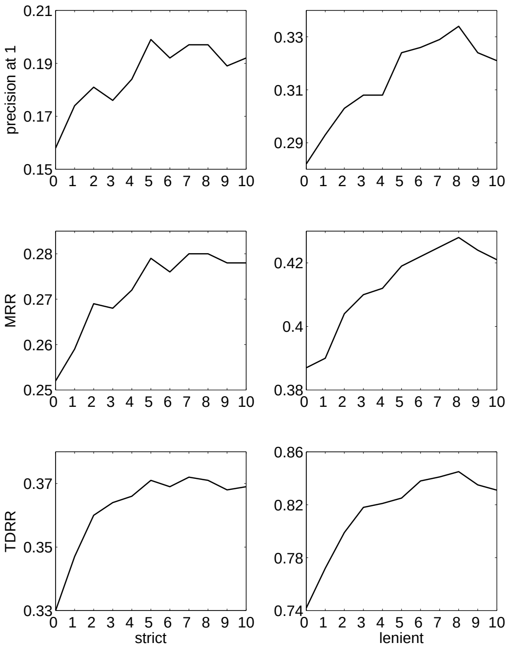

Fig. 1: *Performance of passage retrieval for different expansion lengths corresponding to k values from 1 to 10 under strict and lenient criteria.*

5 Conclusion

In a Question Answering system, an answer to a question cannot be retrieved unless it is present in one of the retrieved passages. So, passage retrieval is considered as one of the most important components of a QA system. Techniques like query expansion are often used to improve the performance of information retrieval systems. In this paper, we have described a new query expansion method which aims to rank the answer containing passages better. It used content and structured information (link structure and category information) of Wikipedia to generate a set of semantically related terms to the question. An empirical evaluation using TREC 2006 QA data set showed significant improvements using our query expansion method.

References

[1] J. Arguello, J. Elsas, J. Callan, and J. Carbonell. Document representation and query expansion models for blog recommendation. In *Int. Conf. on Weblogs and Social Media (ICWSM)*, 2008.

[2] K. B. Bilotti, M.W. and J. Lin. What works better for question answering: Stemming or morphological query expansion? In *ACM SIGIR'04 Workshop Information Retrieval for QA*, 2004.

[3] H. T. Dang, J. J. Lin, and D. Kelly. Overview of the trec 2006 question answering track 99. In *TREC*, 2006.

[4] L. Derczynski, J. Wang, R. Gaizauskas, and M. A. Greenwood. A data driven approach to query expansion in question answering. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008) Workshop on Information Retrieval for Question Answering (IR4QA08)*, 2008.

[5] D. Moldovan, M. Paca, S. Harabagiu, A. Harabagiu, and M. Surdeanu. Performance issues and error analysis in an open-domain question answering system. In *ACM Trans. Inf. Syst*, page 2003, 2002.

[6] C. Monz. From document retrieval to question answering. In *ILLC Dissertation Series*, 2003.

[7] L. Pizzato, D. Mollá, and C. Paris. Pseudo relevance feedback using named entities for question answering. In *Proceedings ALTW*, volume 4, pages 83–90, 2006.

[8] R. Sun, C.-H. Ong, and T.-S. Chua. Mining dependency relations for query expansion in passage retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 382–389, New York, NY, USA, 2006. ACM.

[9] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47, New York, NY, USA, 2003. ACM.

[10] H. Yang, T.-S. Chua, S. Wang, and C.-K. Koh. Structured use of external knowledge for event-based open domain question answering. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 33–40, New York, NY, USA, 2003. ACM.