Exploration of Term Dependence in Sentence Retrieval

Keke Cai, Jiajun Bu, Chun Chen, Kangmiao Liu
College of Computer Science, Zhejiang University
Hangzhou, 310027, China
{caikeke,bjj,chenc,lkm}@zju.edu.cn

Abstract

This paper focuses on the exploration of term dependence in the application of sentence retrieval. The adjacent terms appearing in query are assumed to be related with each other. These assumed dependences among query terms will be further validated for each sentence and sentences, which present strong syntactic relationship among query terms, are considered more relevant. Experimental results have fully demonstrated the promising of the proposed models in improving sentence retrieval effectiveness.

1 Introduction

Sentence retrieval is to retrieve sentences in response to certain requirements. It has been widely applied in many tasks, such as passage retrieval (Salton et al, 1994), document summarization (Daumé and Marcu, 2006), question answering (Li, 2003) and novelty detection (Li and Croft 2005). A lot of different approaches have been proposed for this service, but most of them are based on term matching. Compared with document, sentence always consists of fewer terms. Limited information contained in sentence makes it quite difficult to implement such term based matching approaches.

Term dependence, which means that the presence or absence of one set of terms provides information about the probabilities of the presence or absence of another set of terms, has been widely accepted in recent studies of information retrieval. Taking into account the limited information about term distribution in sentence, the necessary of incorporating term dependence into sentence retrieval is clear.

Two kinds of dependence can be considered in the service of sentence retrieval. The first one occurs among query or sentence terms and another one occurs between query and sentence terms. This paper mainly focuses on the first kind of dependence and correspondingly proposes a new sentence retrieval model (TDSR). In general, TDSR model can be achieved through the following two steps:

The first step is to simulate the dependences among query terms and then represent query as a set of term combinations, terms of each of which are considered to be dependent with each other.

The second step is to measure the relevance of each sentence by considering the syntactic relationship of terms in each term combination formed above and then sort sentences according to their relevance to the given query.

The remainder is structured as follows: Section 2 introduces some related studies. Section 3 describes the proposed sentence retrieval model. In Section 4, the experimental results are presented and section 5 concludes the paper.

2 Related Works

Sentence retrieval is always treated as a special type of document retrieval (Larkey et al, 2002; Schiffman, 2002; Zhang et al, 2003). Weight function, such as tfidf algorithm, is used to construct the weighted term vectors of query and sentence. Similarity of these two vectors is then used as the evidence of sentence relevance. In fact, document retrieval differs from sentence retrieval in many ways. Thus, traditional docu-

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 97–100,
Prague, June 2007.

ment retrieval approaches, when implemented in the service of sentence retrieval, cannot achieve the expected retrieval performance.

Some systems try to utilize linguistic or other features of sentences to facilitate the detection of sentence relevance. In the study of White (2005), factors used for ranking sentences include the position of sentence in the source document, the words contained in sentence and the number of query terms contained in sentence. In another study (Collins-Thompson et al., 2002), semantic and lexical features are extracted from the initial retrieved sentences to filter out possible non-relevant sentences. Li and Croft (2005) chooses to describe a query by patterns that include both query words and required answer types. These patterns are then used to retrieve sentences.

Term dependence also has been tried in some sentence retrieval models. Most of these approaches realize it by referring to query expansion or relevance feedback. Terms that are semantically equivalent to the query terms or co-occurred with the query terms frequently can be selected as expanded terms (Schiffman, 2002). Moreover, query also can be expanded by using concept groups (Ohgaya et al., 2003). Sentences are then ranked by the cosine similarity between the expanded query vector and sentence vector. In (Zhang et al., 2003), blind relevance feedback and automatic sentence categorization based Support Vector Machine (SVM) are combined together to finish the task of sentence retrieval. In recent study, a translation model is proposed for monolingual sentence retrieval (Murdock and Croft, 2005). The basic idea is to use explicit relationships between terms to evaluate the translation probability between query and sentence. Although the translation makes an effective utilization of term relationships in the service of sentence retrieval, the most difficulty is how to construct the parallel corpus used for term translation.

Studies above have shown the positive effects of term dependence on sentence retrieval. However, it is considered that for the special task of sentence retrieval the potentialities of term dependence have not been fully explored. Sentence, being an integrated information unit, always has special syntactic structure. This kind of information is considered quite important to sentence relevance. How to incorporate this kind of information with information about dependences in query to realize the most efficient sentence retrieval is the main objective of this paper.

3 TDSR Model

As discussed above, the implementation of TDSR model consists of two steps. The following will give the detail description of each step.

3.1 Term Dependences in Query

Past studies have shown the importance of dependences among query terms and different approaches have been proposed to define the styles of term dependence in query. In this paper, the assumption of term dependence starts by considering the possible syntactic relationships of terms. For that the syntactic relationships can happen among any set of query terms, hence the assumption of dependence occurring among any query terms is considered more reasonable.

The dependences among all query terms will be defined in this paper. Based on this definition, the given query Q can be represented as: $Q = \{TS_1, TS_2, \ldots, TS_n\}$, each item of which contains one or more query terms. These assumed dependences will be further evaluated in each retrieved sentence and then used to define the relevance of sentence

3.2 Identification of Sentence Relevance

Term dependences defined above provide structure basis for sentence relevance estimate. However, their effects to sentence relevance identification are finally decided by the definition of sentence feature function. Sentence feature function is used to estimate the importance of the estimated dependences and then decides the relevance of each retrieved sentence.

In this paper, feature function is defined from the perspective of syntactic relationship of terms in sentence. The specific dependency grammar is used to describe such relationship in the form of dependency parse tree. A dependency syntactic relationship is an asymmetric relationship between a word called governor and another word called modifier. In this paper, MINIPAR is adopted as the dependency parser. An example of a dependency parse tree parsed by MINIPAR is shown in Figure 1, in which nodes are labeled by part of speeches and edges are labeled by relation types.

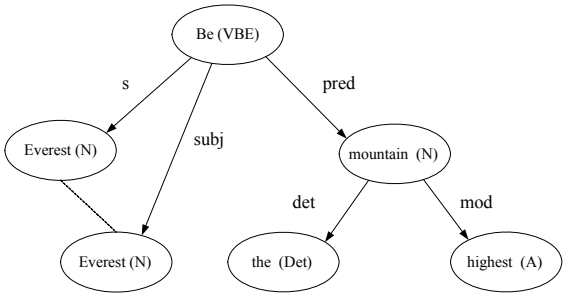

Figure 1. Dependency parse tree of sentence "Everest is the highest mountain".

As we know, terms within a sentence can be described by certain syntactic relationship (direct or indirect). Moreover, different syntactic relationships describe different degrees of associations. Given a query, the relevance of each sentence is considered different if query terms present different forms of syntactic relationships. This paper makes an investigation of syntactic relationships among terms and then proposes a novel feature function.

To evaluate the syntactic relationship of terms, the concept of association strength should be defined to each $TS_i \in Q$ with respect to each sentence S. It describes the association of terms in TS_i. The more closely they are related, the higher the value is. In this paper, the association strength of TS_i is valued from two aspects:

- Size of TS_i. Sentences containing more query terms are considered more relevant.
- Distance of TS_i. In the context of dependency parse tree, the link between two terms means their direct syntactic relationship. For terms with no direct linkage, their syntactic relationship can be described by the path between their corresponding nodes in tree. For example, in Figure 1 the syntactic relationship between terms "Everest" and "mountain" can be described by the path:

$$\text{Everest} \xrightarrow{\text{s}} \text{Be} \xrightarrow{\text{pred}} \text{mountain}$$

This paper uses term distance to evaluate terms syntactic relationship. Given two terms A and B, their distance $distance(A, B)$ is defined as the number of linkages between A and B with no consideration of direction. Furthermore, for the term set C, their distance is defined as:

$$D(C) = \frac{1}{N} * \sum_{q_i, q_j \in C} distance(q_i, q_j) \quad (1)$$

where N is the number of term pairs of C.

Given the term set TS_i, the association strength of TS_i in sentence S is defined as:

$$AS(TS_i, S) = \alpha^{1/S(TS_i)} * \beta^{D(TS_i)} \quad (2)$$

where $S(TS_i)$ is the size of term set TS_i and parameters α and β are valued between 0 and 1 and used to control the influence of each component on the computation of $AS(TS_i)$.

Based on the definition of association strength, the feature function of S can be further defined as:

$$F(S, Q) = \max_{TS_i \in Q} AS(TS_i, S) \quad (3)$$

Taking the maximum association strength to evaluate sentence relevance conforms to the Disjunctive Relevance Decision principle (Kong et al., 2004). Based on the feature function defined above, sentences can be finally ranked according to the obtained maximum association strength.

4 Experiments

In this paper, the proposed method is evaluated on the data collection used in TREC novelty track 2003 and 2004 with the topics N1-N50 and N51-N100. Only the title portion of these TREC topics is considered.

To measure the performance of the suggested retrieval model, three traditional sentence retrieval models are also performed, i.e., TFIDF model (TFIDF), Okapi model (OKAPI) and KL-divergence model with Dirichlet smoothing (KLD). The result of TFIDF provides the baseline from which to compare other retrieval models.

Table 1 shows the non-interpolated average precision of each different retrieval models. The value in parentheses is the improvement over the baseline method. As shown in the table, TDSR model outperforms TFIDF model obviously. The improvements are respectively 15.3% and 10.2%.

	N1-N50	**N51-N100**
TFIDF	0.308	0.215
OKAPI	0.239 (-22.4)	0.165 (-23.3%)
KLD	0.281 (-8.8)	0.204 (-5.1%)
TDSR	0.355 (15.3%)	0.237 (10.2%)

Table 1. Average precision of each different retrieval models

Figure 2 and Figure 3 further depict the precision recall curve of each retrieval model when implemented on different query sets. The improvements of the proposed retrieval model indicated in these figures are clear. TDSR outperforms other retrieval models at any recall point.

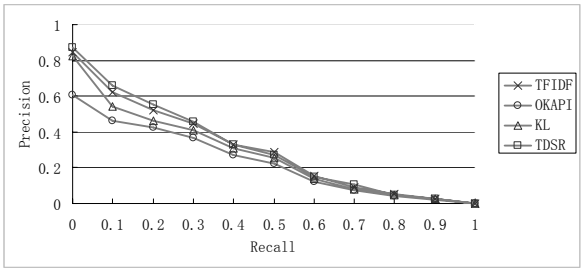

Figure 2. Precision-Recall Curve of Each Retrieval Model (N1-N50)

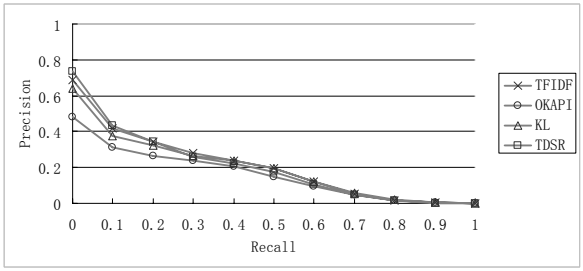

Figure 3. Precision-Recall Curve of Each Retrieval Model (N51-N100)

5 Conclusions

This paper presents a novel approach for sentence retrieval. Given a sentence, its relevance is measured by the degree of its support to the dependences between query terms. Term dependence, which has been widely considered in the studies of document retrieval, is the basis of this retrieval model. Experimental results show the promising of the proposed models in improving sentence retrieval performance.

References

Barry Schiffman. 2002. Experiments in Novelty Detection at Columbia University. In *Proceedings of the 11th Text REtrieval Conference*, pages 188-196.

Gerard Salton, James Allan, and Chris Buckley. 1994. Automatic structuring and retrieval of large text files. *Communication of the ACM*, 37(2): 97-108.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 305-312, Sydney, Australia.

Kevyn Collins-Thompson, Paul Ogilvie, Yi Zhang, and Jamie Callan. 2002. Information filtering, Novelty detection, and named-page finding. In *Proceedings of the 11th Text REtrieval Conference*, National Institute of Standards and Technology.

Leah S. Larkey, James Allan, Margaret E. Connell, Alvaro Bolivar, and Courtney Wade. 2002. UMass at TREC 2002: Cross Language and Novelty Tracks. In *Proceeding of the Eleventh Text Retrieval Conference*, pages 721–732, Gaithersburg, Maryland.

Min Zhang, Chuan Lin, Yiqun Liu, Le Zhao, Liang Ma, and Shaoping Ma. 2003. THUIR at TREC 2003: Novelty, Robust, Web and HARD. In *Proceedings of 12th Text Retrieval Conference*, pages 137-148.

Ryen W. White, Joemon M. Jose, and Ian Ruthven. 2005. Using top-ranking sentences to facilitate effective information access. *Journal of the American Society for Information Science and Technology*, 56(10): 1113-1125.

Ryosuke Ohgaya, Akiyoshi Shimmura, Tomohiro Takagi, and Akiko N. Aizawa. 2003. Meiji University web and novelty track experiments at TREC 2003. In *Proceedings of the Twelfth Text Retrieval Conference*.

Vanessa Murdock and W. Bruce Croft. 2005. A translation Model for Sentence retrieval. HLT/EMNLP. In *Proceedings of the Conference on Human Language Technologies and Empirical Methods in Natural Language Processing*, pages 684-691.

Xiaoyan Li. 2003. Syntactic Features in Question Answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 455-456, Toronto, Canada.

Xiaoyan Li and W. Bruce Croft. 2005. Novelty detection based on sentence level patterns. In *Proceedings of ACM Fourteenth Conference on Information and Knowledge Management (CIKM)*, pages 744-751, Bremen, Germany.

Y.K. Kong, R.W.P. Luk, W. Lam, K.S. Ho and F.L. Chung. 2004. Passage-based retrieval based on parameterized fuzzy operators, *ACM SIGIR Workshop on Mathematical/Formal Methods for Information Retrieval*.