Exploiting Question Concepts for Query Expansion

Hae-Jung Kim, Ki-Dong Bu, Junghyun Kim, and Sang-Jo Lee

Department of Computer Engineering, Kyungpook National University,
Sangyuk-dong, Puk-gu, Daegu, 702-701, Korea
hjkim325@hanmail.net

Abstract. In this paper, we present an efficient semantic query expansion methodology based on a question concept list comprised of terms that are semantically close to concepts represented in a query. The proposed system first constructs a concept list for each question concept and then learns the concept list for each question concept. When a new query is given, the question is classified into the question concept, and the query is expanded using the concept list of the classified concept. In the question answering experiments on 42,654 Wall Street Journal documents of the TREC collection, the traditional system showed in 0.223 in MRR and the proposed system showed 0.50 superior to the traditional question answering system.

1 Introduction

Question answering (QA) systems assign relevance degrees to words, paragraphs or clauses based on a given query, and then provide answers ranked according to relevance. However, the efficacy of such systems is limited by the fact that the terms used in a query may be in a syntactic form different to that of the same words in a document. Consider, for example, the following query and sentences:

- Who is the inventor of a paper?
- S1: C is the inventor of knives
- S2: a devised paper in China…

When analyzing this query, the traditional QA system would classify the sample query into "NAME" as a subcategory of "PERSON", and then keywords such as "inventor" and "paper" would be extracted. In this example, however, S1 contains the keyword "inventor" and S2 contains the keyword "paper", and hence their relevance degrees for the query will be the same. Moreover, even if we expand the keywords to "inventor", "discoverer", and "paper", the ranking of the sample sentences will remain unchanged because the term "devise" in S2 belongs to a syntactic category different to that of "inventor" in the query. However, if we were to expand the keyword "inventor" to include related words such as "discoverer", "devise", "invent", "develop", and "creator", then we could represent the same concept over a range of syntactic and semantic categories, and thereby reduce the number of answer candidates and extract more exact answers.

In this paper, we present an efficient semantic query expansion methodology based on a question concept list comprised of terms that are semantically close to concepts

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 624 – 627, 2005.

represented in a query. The concept list associated with a particular query includes most possible representations of the concept of the question.

2 Previous Work

Answer type of QA system can be called semantic category of the query that a user requested, and it had an influence on a QA system performance enhancement to express answer type as the small classification of semantic category [1-4]. Cardie *et al.* [1] modified the traditional approach to question type classification by dividing the answer type into 13 subcategories, thereby creating more specific question categories. This modification significantly improved the performance of the traditional QA system. Prager *et al.* [4] proposed an alternative methodology for finding the semantic class that covering all possible semantic classes used in a query; specifically, they determined a synset of question terms by using an inventory such as a hypernym tree from WordNet. However, their method entails the derivation of the synset-class mapping, which is a labor-intensive task that results in incomplete coverage. In contrast to the above methods, our method contains the concept list of each question concept that can be used to expand query terms into conceptually close terms.

3 Query Expansion Based on a Question Concept List

3.1 System Description

Figure 1 shows a flow diagram of the overall system configuration. The system contains three main components: In the question concept list construction module, the concept list of each question concept is constructed for query expansion. First, the concepts of question categories are established according to important question concepts by the question concept classification module. Then, the concept list for each question concept is constructed by query pattern recognition. In the concept pattern learning module, the system learns the constructed concept list of each question concept using a learning algorithm. Finally, in the question analysis component, the system classifies the given query into the corresponding question concept node based on learned data, and then expands the query into the semantically close terms using the concept list of a classified concept node.

3.2 Question Concept List Construction

We assume that the important concept of a question will be embodied in the terms that are most frequently used in the question; hence, we categorize the question type based on the term frequency (TF) of two categories, nouns and verbs. We regard terms occupying the upper 30% of the total TF, and the question concepts of 117 Who queries from TREC-9 collection can be categorized into 8 concepts such as “inventor, killer, writer, leader, player, founder, owner, others”.

To construct the concept list of each question concept, we should extract the terms that represent the concepts of each question. To facilitate extraction of the concept of the query, we extract the pattern of the query as defined in Definition 1. For example, the question pattern from the query "Who is the inventor of paper?" is <Who, null, is_BE, inventor_NN>.

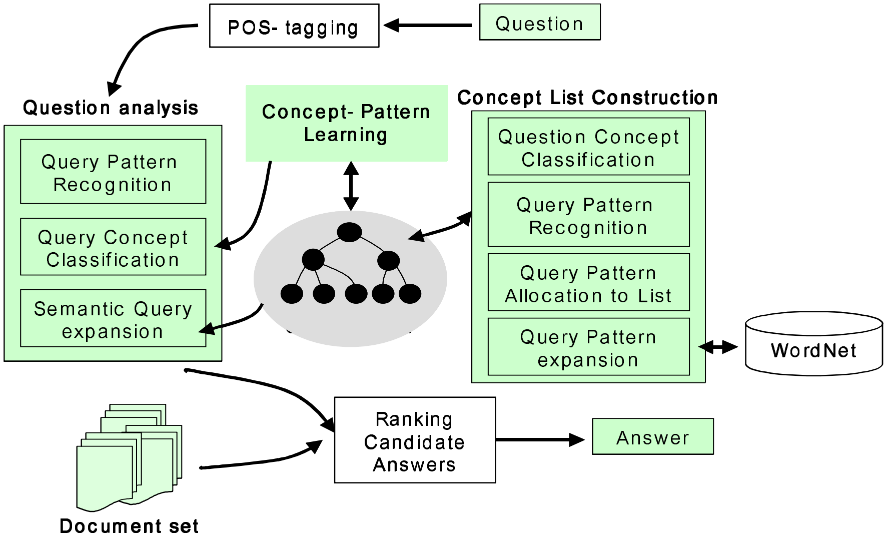
POS- tagging
Question
Question analysis
Query Pattern Recognition
Query Concept Classification
Semantic Query expansion
Concept- Pattern Learning
Concept List Construction
Question Concept Classification
Query Pattern Recognition
Query Pattern Allocation to List
Query Pattern expansion
WordNet
Ranking Candidate Answers
Answer
Document set

Fig. 1. Overall System Configuration

[Definition 1: Question pattern]
Question patterns are defined as the following two types based on the noun (N) and verbs (BE_V, V) around Wh_term, where, BE_V is the verb "to be" or one of its conjugated forms. Noun N1 is the first noun before verb V and noun N2 is the first noun after verb V.

Question pattern 1 = [Wh_term, N1, BE_V, N2]
Question pattern 2 = [Wh_term, N, V]

The patterns extracted from a query are assigned to the corresponding question concept and make up the "concept list" that represents the concept of a question.

3.3 Question Concept Learning and Query Expansion

For concept learning and classification, we use the Naïve Bayes theorem. For the terms v_j in a pattern, we calculate the $P(v_j|S_k)$ and $P(v_j)$ for all concepts k and select the S_k that has the highest probability as the question concept S' of the given query.

$$Decide\ \ S'\ \ if\ \ \ S' = \arg\max_{S_k} [\log P(S_k) + \sum_{v_j\ in\ C} \log P(v_j \mid S_k)]$$

For a new query, the proposed system extracts the query pattern and then classifies the query into the question concept based on the learned data. The system then acquires the expansion terms from the concept list in the classified question concept.

4 Evaluation

To test the proposed method, we first tested the classification performance of the constructed question concept list, and applied the proposed query expansion method to the question answering system. We used precision as measures of the accuracy.

When the total number of patterns in the constructed concept list was 117 Who queries from TREC-9 collection learning and classification performance were 94.4% precision for the learning set and 78.7% for the test set by 10-fold cross validation.

Table 3. Precision in 10-fold cross validation for concept list learning

	Traning set	Test set
Micro avg.	0.944	0.787

We conducted retrieval test on the Wall Street Journal (WSJ) 1991 with 42,654 documents and 18 who-queries in TREC-9 collection having WSJ 1991 documents as answer set. Similarity measure between questions and documents was:

$$\mathrm{Sim}(Q, D) = \sum_{\mathrm{i}}\sum_{\mathrm{j}}\alpha_I \times \delta(q_i, d_j), \quad \text{where } \delta(q_i, d_j) = 1 \text{ if } q_i = d_j, \text{ otherwise } 0.$$

Table 4 shows the Mean Reciprocal Ratio (MRR) results for the comparison of the traditional QA system and the proposed. The proposed system showed 0.50 superior to the traditional system when the sentence boundary was three sentences.

Table 4. MRR of the traditional QA system and the proposed system

	Traditional	The proposed
Three sentences	0.223	0.500

5 Conclusions

In this paper, by assuming that the important concepts of a query are embodied in the most frequently used terms in the query, we constructed a question concept list that contains an expanded collection of query terms related to the concept of a query. When we evaluated the performance of the proposed method, the proposed system showed 0.50 in MRR superior to the traditional system. The results of the present experiments suggest the promise of the proposed method.

References

1. C., Cardie, V., Ng, D., Pierce and C., Buckley, Examining the Role of Statistical and Linguistic Knowledge Sources in a General-Knowledge Question-Answering System, In Proceeding of the 6th Applied Natural Language Processing Conference, 2000, 180-187
2. E. Hovy, et al., Learning Surface Text Patterns for a Question Answering system, In Proceedings of the ACL conference, 2002, 180-187.
3. H. Kazawa, T. Hirao, H. Isozaki, and E. Maeda, A machine learning approach for QA and Novelty Tracks:NTT system description, In Procs. of the 11th Text Retrieval Conf., 2003.
4. J. Prager, D. Radev, E. Brown, A. Coden, The Use of Predictive Annotation for Question-Answering in TREC8, In Proceedings of the TREC-8 Conference NIST, 2000, 309-316