THUIR at TREC 2004: QA

Wei Tan, Qunxiu Chen, Shaoping Ma
State Key Lab of Intelligent Tech. & Sys., CS&T Dept, Tsinghua University, Beijing 100084, China
tw02@mails.tsinghua.edu.cn

Abstract

In this paper, we describe ideas and related experiments of Tsinghua University IR group in TREC 2004 QA track. In this track, our system consists three components: Question analysis, Information retrieval, and Answer extraction. Question analysis component extracts Query Term and answer type. Information retrieval component retrieves candidate documents from index set based on paragraph level and re-ranks them to find more relevant documents. And then Answer extraction component matches empirical phrases according to answer type to extract final answer.

1. Introduction

TREC introduced the first question answering(QA) track in TREC-8(1999).The goal of the track is to foster research on systems that retrieve answers rather than documents in response to a question, with particular emphasis on systems that can function in unrestricted domains. The tasks in the track have evolved over the years to focus research on particular aspects of the problem deemed important to improving the state-of-the-art.

Compared to TREC 2003, TREC 2004 QA track has done some changes: The TREC 2004 QA track consists of a single task that is a combination of factoid, list, and definition-like questions. There are 65 targets of a definition. For each target, there are a series of factoid and list questions that relate to that target. The final question in each series is an explicit "other" question that should be interpreted as "tell me other interesting things about this target I didn't know enough to ask directly". This final question is roughly equivalent to the TREC 2003 QA track's definition question.

Section 2 describes the overview of THUIR QA system. The system consists of three components: Question analysis, Information retrieval, Answer extraction, as detailed in section 3, 4 and 5. Section 6 presents the final result and evaluation. Finally, section 7 describes the discussion and future work.

2. Overview of THUIR Question Answering System

Similar to other systems, our system consists of three components: Question analysis, Information retrieval, and Answer extraction. The architecture is illustrated by Figure 2.1.

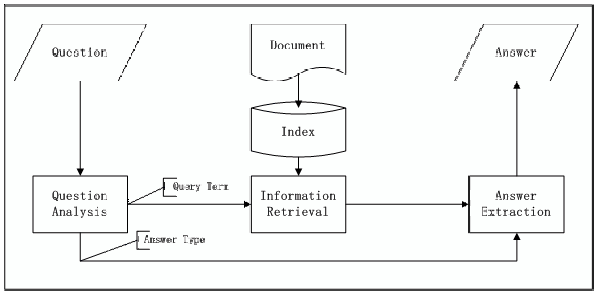

Figure 2.1: Architecture of the THUIR QA System

The Question Analysis component processes user questions and extracts useful information: Query Term and Answer Type. Ranked candidate documents are retrieved by Information Retrieval component from indexed documents set. Each candidate document is assigned with a rank score according to the similarity between candidate document and the Query Term. Then the top-ranked documents are processed by Answer Extraction component, and the final answer can be extracted with the Answer Type and some empirical feature matching strategy.

3. Question Analysis

In contrast with several years ago, questions in TREC 2004 QA track are not independent, but context-sensitive, such as there are 65 targets, and each target consist a topic and several questions correlated to the topic. In some questions, some topic words are replaced by pronoun with the context of the question and the topic. So, it is necessary to consider of the relationship between the pronouns of questions and the topic words when questions are analyzed. In our system, pronouns reference to the topic are replaced by topic words and extracted as a Query Term.

Answer Type of questions is classified according to the interrogative word and the sentence structure, listed in Table 3.1.

Answer Type	Sub-Answer Type	Question Structure
PERSON	Name	who; what/which + occupation name; …
	Position	who + is/was + person name; …
NUMBER	Duration	how long; how many years; how many days; …
	Age	how old; …
	Length	how long; how high; how far; how close; How tall; …
	Area	how big; how large; …
	Volume	how big; …
	Size	how big; what … size; …
	Weight	how big; …
	Rate	how much; what percent; …
	Money	how much; …
	Frequency	how often; how fast; …
	Temperature	how hot; how cold; …
	Quantity	how many + plural noun; …
MATTER	Definition	what is/was …
	Object	what do/does/did … VP
	What+NP	what NP …
TIME	Time	when; what date; what day; …
PLACE	Place	where; what country; what city; …
METHOD	Method	how
REASON	Reason	why
IMPERATIVE	Imperative	name; list; …

Table 3.1: Answer Type of question

TIME, PLACE, METHOD and REASON type are easy to be classified by interrogative word, so these types are not subdivided. Imperative sentences are generally treated as MATTER type. Because the questions just as “who is person-name” type are considered to ask the position

of the "person-name", PERSON type need to be subdivided into Name and Position sub-type. Questions like "how + adj/adv" type are mostly answered by number, and then with the difference of adj/adv NUMBER type is subdivided into several sub-type such as duration, age, size, money, and so on. Among the questions beginning with "what", questions as "what is/was NP" type are assigned to Definition sub-type, questions in which "what" is object are assigned to Object sub-type, and questions as "what + NP" type is assigned to What+NP sub-type.

NP chunk is extracted and NE (Named Entity) is recognized by GATE from questions, and every elements of question are experientially assigned a weight such as NE and number is 10, noun is 5, notional verb is 3, adjective and adverb is 1, and auxiliary verb, interrogative word and punctuation is 0. In addition, the topic word of the target, which questions belong to, is treated as Query Term and assigned weight 5.

4. Information Retrieval

The document set uses documents on the AQUAINT disk set just as the past two QA tracks. Each document is divided several paragraphs, and the index is built to each paragraph. Furthermore, another index is built to "Headline" of each document for answer extraction of questions of "Other" type. Index built and information retrieval uses the Lemur of CMU, retrieval model is simple tfidf model, and top-ranked 1000 documents are candidate documents.

And then candidate documents are re-ranked by the formula given below:

$$Score_{rerank} = \alpha \cdot Score_{qt} + \beta \cdot Score_{NNP} + \gamma \cdot Score_{d}; \tag{4.1}$$

$$Score_{qt} = \frac{QTCount_{doc}}{QTCount_{q}}; \tag{4.2}$$

$$Score_{NNP} = \begin{cases} 0 & if\ no\ NNP \\ \dfrac{NNPCount_{doc}}{NNPCount_{q}} & Otherwise \end{cases} \tag{4.3}$$

$$Score_{d} = \begin{cases} 0 & if\ number\ of\ existing\ query\ term\ is\ not\ more\ than\ 1 \\ 1 - \sqrt{\dfrac{\sum_{i=1}^{n}(POS_i - POS_{mean})^2}{(n-1) * Length_{doc}^{\ 2}}} & Otherwise \end{cases} \tag{4.4}$$

Where, $Score_{rerank}$ is the score of documents re-ranked, and it consists three parts: $Score_{qt}$, $Score_{NNP}$, and $Score_{d}$. $Score_{qt}$ presents the rate of the number of Query Term in documents and total one. $QTCount_{q}$ is total number of Query Term, and $QTCount_{doc}$ is the number of the Query Term existing in documents. If existing, the Query Term counts once. The rate of proper nouns extracted from questions is presented by $Score_{NNP}$. If no proper nouns, $Score_{NNP}$ is 0; otherwise, $Score_{NNP}$ is the rate of $NNPCount_{doc}$, the number of proper nouns in documents,

and $NNPCount_q$, total number of proper nouns, and just as $QTCount_{doc}$, if a proper noun exist in documents, it only counts once. $Score_d$ is the density of Query Term distributing in documents. If the number of Query Term in documents is no more than 1, $Score_d$ is 0. n is the total number of Query Term in documents, POS_i is the position of the *i*th Query Term in documents, POS_{mean} is the mean of all Query Terms, and $Length_{doc}$ is the length of the document.

From the definition of $Score_{qt}$, $Score_{NNP}$, and $Score_d$, the three score is just between 0 and 1. The more the Query Terms appear in documents, the closer $Score_{qt}$ is to 1, and then we consider these documents are more relevant to questions. $Score_{NNP}$ is just same as $Score_{qt}$. When all query terms in documents centralize, the $Score_d$ is close to 1. We think that the answer may be extracted from the sentence in which more query terms centralize. So we use the linear sum of these three score as re-rank score. In our experiments, linear coefficients are all 1/3. After re-ranked, the more relevant documents can be ranked ahead.

5. Answer Extraction

First of all, candidate documents are processed by a named entity finder (GATE), which recognizes NE nouns such as person name, location, organization, date, time, money, etc. Different strategy of answer extraction is adapted to different answer type.

1) For answer type as PERSON-name, TIME, and PLACE, answer can be extracted from corresponding named entity nouns.
2) For NUMBER type, sub-type decides the answer extraction method. For "age" sub-type, phrase "number + years + old" is matched in documents; For "duration" sub-type, phrase "number + time noun" is expected as an answer; Other sub-type is just similarly processed to match some empirical phrase.
3) MATTER type is processed as NUMBER type. NP_d is noun phrase to be defined in question of "definition" sub-type, and then phrase "NP_d + be", "NP_d, known as", etc is matched in documents. For "object" sub-type, subject noun phrase NP_s and verb phrase VP is extracted from questions, phrase "$NP_s + VP$" is matched in documents, and then the object of matched sentence is expected as answer. For "what+NP" sub-type, phrase "NP + VP" is matched, where VP is verb phrase of questions. The modifier of NP and NP is extracted as an answer.

If there are several matched phrases, each phrase is assigned a weight score according to the distance between the phrase position and central position of all Query term in documents.

6. Evaluation

In TREC2004 QA track, we submitted one run, and here is evaluation result:

	factoid		list	other	final score
	Precision	Recall	F Score	F Score	
THUQA04RUN1	0.091	0.318	0.085	0.055	0.085

Table 6.1: The TREC Evaluation Result

7. Discussion and Future Work

This is the first time that Tsinghua University IR group (THUIR) participates in TREC QA track. The Evaluation shows that our system and method need be improved. We still have a lot of things to do in the future. Question analysis component needs not only syntax analysis but semantic one of questions to a certain extent. Re-rank strategy of Information retrieval component will be mended so that the more relevant documents can be ranked more ahead. Answer extraction component exists very great shortcoming because of empirical phrase extraction strategy. For some questions, the top-ranked documents contain answers, however our answer extraction strategy can not find them. So, we need to research more in answer extraction and build new extraction strategy in the future.

Reference

1. Ellen, Voorhees. Overview of the TREC 2003 Question Answering Track. In TREC 2003, in Gaithersburg, Maryland. NIST
2. J. Prager, D. Radev, E. Brown, A. Coden. The Use of Predictive Annotation for Question Answering in TREC8. In proceedings of the Eighth Text Retrieval Conference, 1999.
3. L. Wu, X-j Huang, Y. Guo, B. Liu, Y. Zhang. FDU at TREC-9: CLIR, Filtering and QA Tasks. In proceedings of the ninth Text Retrieval Conference, 2000.
4. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.