Description of Preliminary Results to TREC-8 QA Task

Chuan-Jie Lin and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
E-mail: cjlin@nlg2.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw

1. Introduction

Question Answering (QA) becomes a hot research topic in recent years due to the very large virtual database on the Internet. QA is defined to find the exact answer, which can meet the users' need more precisely, from a huge unstructured database. Traditional information retrieval systems cannot afford to resolve this problem. On the one hand, users have to find out the answers by themselves from the documents returned by IR systems. On the other hand, the answers may appear in any documents, even that the document is irrelevant to the question.

Two possible approaches, i.e., keyword matching and template extraction, can be considered. Keyword matching postulates that the answering text contains most of the keywords. In other words, it carries enough information relevant to the question. Using templates is some sort of information extraction. The contents of documents are represented as templates. To answer a question, a QA system has to select an appropriate template, then fill the template and finally offer the answer. The major difficulties in this approach are to find general domain templates, and to decide which template can be applied to answer the question.

Some other techniques are also useful. For example, to answer the questions "Who..." and "When...", the identification of named entities like person names and time/date expressions will help to locate the answer.

In our preliminary study, we adopt keyword-matching strategy coupling with expanding the keyword set selected from the question sentence by the synonyms and the morphological forms. We participate in the group “Sentence or under 250 bytes.” The detail will be presented below.

2. Description of Our System

The system is composed of three major steps: (1) preprocessing the question sentences, (2) retrieving the documents containing answers, and (3) retrieving the sentences containing answers.

2.1 Preprocessing the Question Sentences

Our main strategy is keyword matching. This approach has a drawback, i.e., the words used in the question sentences and in the sentences containing the answers may be different. For example, verbs can be in different tenses and synonyms can also be used. Therefore we have to make necessary changes and expansions in the question sentences.

At first the parts-of-speech are assigned to the words in question sentences. Then, stop-words are removed. The remaining words are transformed into the canonical forms and selected as the keywords of the question sentences. For each keyword, we find all of its synonyms from WordNet 1.6. Those terms form an expansion set for the keyword. If the keyword is a noun, a verb, an adjective, or an adverb, all the possible morphological forms of the words in the expansion set are also added into this set. Here the morphological forms are the plural of a noun, different tenses of a verb, and the comparison of an adjective or an adverb. They are shown as follows:

noun AAA: AAAs | AA[s,z,sh]es

verb BBB: BBBed BBBing | BB[e]d BB[e]ing / BBBs | BB[s,z,sh]es

adjective or adverb CCC: CCCer CCCest | CC(y)ier CC(y)iest

The irregular nouns and verbs can be transformed by looking up the WordNet.

2.2 Retrieving the Documents Containing Answers

We implement a full text retrieval system to find the documents that may contain the answers. The purpose is to decrease the number of documents we have to search the answering sentences. Each keyword of a question sentence is assigned a weight, so are their various morphological forms. Those words tagged as NNP and NNPS, which denote proper nouns, have assigned higher weights. This is because they should be presented in the answer. The weights of added synonyms are less than the keywords. The score of a document is computed as follows:

$$score(D) = \sum_{t \in EX(T), t \text{ in } D} weight(T)$$

where T is one of the keywords, and $EX(T)$ its expansion set.

The document containing one keyword or any words in its expansion set earns a score of its weight. For example, consider the Question 30:

<num>Number: 30

What are the Valdez Principles?

Its keywords are “Valdez” and “Principles”, and the expansion sets are [valdez/valdezes/] [principle/principles/rule/rules/precept/precepts/rationale/rationales], respectively. If a document contains “principles” and “rules”, but no “valdez”, its score is determined by “principles” and “rules” only. The word “principles” gives a higher weight since it is proper noun, and the word “rules” gives a lower weight. The score of the document is the sum of these two weights.

Those documents that have scores no less than the threshold are selected as the answering documents. Threshold is set to the sum of weights of the words in the original question sentence. Note that the removed words have no scores. If no documents have scores greater than the threshold, we assume that no answers can be found for the question.

2.3 Retrieving the Sentences Containing Answers

Finally, we examine each sentence in the documents that may contain the answers. Those sentences that contain most words in the expanded question sentence are retrieved. The top five sentences are regarded as the answers. If there are more than five possible answers, we randomly select five of them. To meet the limit of 250 bytes, we truncate the sentences that exceed the limit. On the contrary, if the answer is shorter than the limit, we concatenate it with the next sentences.

3. Results and Discussions

The system run on the 198 questions provided by Q&A Track of TREC-8. The weights of proper noun keywords are set to 100, and the others are set to 1. Among these 198 questions, 60 have answers. Total 25 of them are correct, and 20 answers are at the top scores. The following shows some examples.

<num> Number: 29

What is the brightest star visible from Earth?

Ans: In the year 296036, Voyager 2 will make its closest approach to Sirius, the brightest star visible from Earth. Deep space is benign, so dust and cosmic rays will erode Voyager 2 extraordinarily slowly. In a billion or more years, Sagan said, "there w

<num> Number: 102

Who is the Voyager project manager?

Ans: Until December, Voyager 2 occasionally will glance at Neptune and dark space to improve the accuracy of observations its cameras and instruments made during the Neptune flyby, said Voyager project manager Norm Haynes. Pictures of empty space let engi

We examine the results of formal runs, and find that the system can be improved from several aspects:

(1) execution speed of the system

Owing to the long time required, 138 questions in the formal run do not have answers. After revising our algorithm and running again, we answer 136 questions. The evaluation is done by us ourselves. Total 62 of them are correct, and 42 answers are at the top scores.

(2) anaphor resolution

The answering sentence may contain pronouns or other anaphors referring the constituents in the previous sentences. We have to find the antecedents. Similarly, date expressions such as "today" have to be substituted by an exact time.

(3) phrasal searching

Phrasal searching is helpful in some kind of questions. For example, to answer the questions

<num> Number: 115
What is Head Start?
<num> Number: 40
Who won the Nobel Peace Prize in 1991?

the key phrases "head start" and "Nobel peace prize" are very useful to find the answers.

(4) question type

It is also helpful to identify possible answering candidates that the question is asking for. For example, the date/time expression is particularly preferred for the questions as "What day ..." or "When ...". For questions asking about "How many people ...," we shall offer a numerical answer.

Systems for name entity extraction in a famous message understanding competition (MUC, 1998) can be employed to provide this information.

(5) related words in different part-of-speech (POS)

We found that many answers are in different POSes from those in the question sentences. For example:

<num>Number: 130
When was Yemen reunified?
Ans: ... on 22 May 1990 , when the *Yemeni* community was reunified...
... the *reunification* of North and South Yemen in 1990...

Therefore, we have to add such related words to get better possibility to find the answer.

(6) time information

If the time is specific in the question, such as:

<num>Number: 32

Who received the Will Rogers Award in 1989?

we have to make sure if the answer contains information happening in the specific time. Time information can be mentioned earlier before the answering sentence, or mentioned in the header as the information of the whole document.

There are also cases requiring more semantic information or world knowledge to find the answers.

(1) additional knowledge

For example, the possessive expression " 's " has many different meanings, depending on the relationship in the expression.

[Number: 7] What debts did Qintex group leave -- Qintex's debt

[Number: 11] President Cleveland's wife -- married 21-year-old Frances Folsom

[Number: 94] Who wrote the song, "Stardust"? -- Hoagie Carmichael's "Stardust,"

We have to know that Cleveland's wife is the one who married him, and Carmichael's "Stardust" means that he wrote this song. In this way, we can find the answer correctly.

Other examples are those phrases expressing the same information, but not using synonyms.

[Number: 14] producer of tungsten -- the biggest supplier of the metal

[Number: 34] Where is the actress, Marion Davies, buried? -- on her mausoleum

[Number: 108] created the Internet browser Mosaic -- Mosaic, developed by

We can see that the words "producer" and "supplier" are not synonyms, nor are "create" and "develop". However, they do offer the same information.

(2) ellipsis

[Number: 104] ... in Marathi, the most commonly spoken local language ,

[Number: 111] The distance in time from Tokyo is ...

The answer to Question 104 in fact mentions "the most commonly spoken language in Bombay", but the location is not shown. So is the answer to Question 111, which indeed mentions "from Tokyo to Niigata".

4. Conclusion and Future Work

We propose a method to answer questions mainly based on keyword matching. The keywords in a question sentence is first selected, then all of their synonyms and morphological variants forms the expansion sets. Appropriate weights are assigned to each keyword.

To look for the sentences containing answers, we first employ a full-text retrieval system to select documents that may contain the answer. Then we examine each sentence in these documents to see if it offers the answer. Our approach answers 136 questions, 62 of them are correct, with 40 at top score.

In the future work, we will try to offer answers according to the types of questions. Besides, we will find the related words of keywords in different POSes, resolve anaphors, match patterns in phrasal level, and find the words missing in the sentences containing answers. Different scoring functions will be investigated to get better performance.

Reference

MUC (1998) *Proceedings of 7th Message Understanding Conference*, http://www.muc.saic. com /proceedings/proceedings_index.html.