RitsQA: Ritsumeikan question answering system used for QAC-1

Jun’ichi FUKUMOTO Tetsuya ENDO Tatsuhiro NIWA
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu-shi, Shiga 525-8577, Japan
{fukumoto, t_endo, t_niwa}@nlp.cs.ritsumei.ac.jp

Abstract

In this paper, we describe RitsQA: Ritsumeikan question answering system. In our QA system, input query sentence is firstly analyzed using question patterns consist of description of NE elements, POS of words and surface expressions. Then, query type and some clue words, which are used for document retrieval, are determined. For answer extraction, our system uses word distance between answer candidate and clue words of retrieved documents.

We have participated Task 1 and Task 2. (it was not formal one because we are a member of QAC task organizer.) The results were not so good than we expected. However, this experience was helpful to develop QAC task.

Keywords: *Question answering, word distance, query pattern analysis.*

1 Introduction

We have started Question answering task called QAC as one task of NTCIR Workshop 3 as a member of QAC task organizer. Participation to QAC will be very helpful for organization of QAC tasks and development of our QA technology.

We have been working on development of Named Entity extraction tool called NExT [7] with Mie university in order to open this NE tool for NLP researcher and other research groups. Applicability of this tool is also our purpose of this participation.

We participated Task 1 and Task 2 of QAC although it was not formal one because we are a member of QAC task organizer. For Task1, we have developed QA system based on word distance for answer extraction. For Task2, we used the same system as a system for Task 1. That is, the best answer in Task 1 will be the answer for Task2.

In this paper, we describe RitsQA: Ritsumeikan question answering system used for QAC-1. Our QA system analyzes input query sentence using question pattern consists of NE elements, POS of words and surface expression. After this analysis, query type and some clue words that are used for document retrieval are determined. For answer extraction, our system is based on word distance from retrieved documents.

2 Query patterns

At the first step of our QA system, query type of input sentences is analyzed. We assumed 5 types of query based on surface expression as follows:

- Who type (“dare” in Japanese)

  This type is a query to get a person name.
- When type (“itsu” in Japanese)

  This type is a query about date or time.
- Where type (“doko” in Japanese)

  This type of a query is used to get information of location name or organization name.
- What type (“nani” in Japanese)

  This type of a query covers a variety of query elements such as person name, organization name, location name, and so on.
- How type (“donokurai” in Japanese)

  This type of a query is used for some numeric expressions such as information of money, ratio (percent), length, weight, speed and so on.

In order to get answers correctly, it is necessary to understand what is required in a query sentence. We have prepared about 70 query patterns to analyze answer type information. A query pattern consists of the following elements.

- Named Entity which is indicated by <PERSON> (person name), <DATE> (date), <ORGANIZATION> (organization name) and so on Table 1 summarizes all the type of Named Entity [1] which are used in our QA system.

[1] As you can see, there NE elements beyond the definition of Named Entity used in IREX [3] and MUC [4]. We used these type of NE elements although all the elements are not tagged in the current implementation of our QA system.

- a word which are indicated by its part of speech (<noun>, <verb> and so on)
- surface expressions

Table 1. NE tags

tag	tag type
<PERSON>	person name
<ARTIFACT>	artifact
<ORGANIZATION>	organization name
<LOCATION>	location name
<DATE>	date
<TIME>	time expression
<MONEY>	money
<PERCENT>	percentage
<UNIT>	unit name
<DISTANCE>	distance expression

Each query pattern has its answer type of information. Example of query patterns and their answer types which are in the brackets shown as follows:

- <noun> は誰 (<PERSON>)
- <verb> のは誰 (<PERSON>)
- <noun> はいつ (<TIME> or <DATE>)
- <verb> のはいつ (<TIME> or <DATE>)
- <noun> は [いつから | いつまで] (<TIME> or <DATE>)
- <noun> が [始まったのは | 終ったのは] いつ (<TIME> or <DATE>)
- <verb> のは [いつから | いつまで] (<TIME> or <DATE>)
- <noun> はどこ (<LOCATION>)
- <verb> のはどこ (<LOCATION>)
- <ARTIFACT> は何 (<ARTIFACT>)
- <PERSON> は何 (<PERSON>)
- <NAME> は何 (<PERSON>)
- <ORGANIZATION> は何 (<ORGANIZATION>)
- <PERSON> の名前は何 (<PERSON>)
- <PERSON> は何という名前 (<PERSON>)
- <NAME> は何という名前 (<PERSON>)
- <ORGANIZATION> は何という名前 (<ORGANIZATION>)
- <LOCATION> は何という名前 (<LOCATION>)
- <ARTIFACT> は何という名前 (<ARTIFACT>)
- <noun> は何という <ARTIFACT> (<ARTIFACT>)
- <noun> は何という <ORGANIZATION> (<ORGANIZATION>)
- <verb> のは何という <PERSON> (<PERSON>)
- <noun> は何 <LOCATION> (<LOCATION>)
- <verb> のは何 <ORGANIZATION> (<ORGANIZATION>)
- <noun> は幾らぐらい (<number>)
- <noun> はどのくらい (<number>)
- <noun> は幾ら (<MONEY>)
- <noun> は何 [円 | ドル..] (<MONEY>)
- <noun> は何 [メートル | グラム | キロ..] (<UNIT>)
- <noun> の [速さ | 長さ | 重さ..] はどのくらい (<UNIT>)
- <noun> はどのくらいの [速さ | 長さ | 重さ..] (<UNIT>)
- <noun> は何% (<PERCENT>)
- <noun> は〜の [前 | 後] 何% (<PERCENT>)
- <noun> から <noun> までどのくらい (<DISTANCE>)

3 Analysis of queries

At first, input query is morphologically analyzed using ChaSen [5] system. Then, NE elements are detected using some simple patterns and clue words for text retrieval are extracted from the query sentence. Clue words that will be extracted are a series of noun, verb, adjective, adverb, number expression and unknown words. NE elements are also used for clue words. After that, answer type is detected by matching the analyzed sentence with query patterns.

4 Text retrieval using query words

We utilized full text search system “NAMAZU”[6] for document retrieval from newspaper database. In our QA system, all clue words extracted from a query sentence are used for Namazu index words. The retrieved documents are order in some probability obtained by NAMAZU system. If there is a clue word that does not match with documents during document retrieval, such a word will be paraphrased to the other expression using thesaurus dictionary. We utilized a small size (about 100 words) dictionary that is made by ourselves.

5 Answer extraction

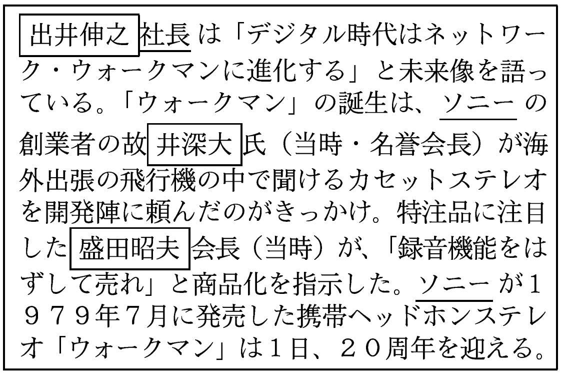
In our QA system, answer extraction is based on word distance between answer candidate and clue words. For each answer candidate, word distance with all clue words are calculated and the sum of the distances will be score of the answer candidate.

At first, top ten documents retrieved by NAMAZU system are morphologically analyzed using ChaSen system and Named Entity elements in the documents are tagged by NExT system [7]. NExt system tags person name, organization name, location name, percentage and money, therefore answer candidates will be limited to these elements in our current implementation.

For each answer candidate who is required answer type, word distance between the candidate and clue words are calculated. For the following query sentence

“ソニーの社長はだれですか。”
(Who is the president of SONY?)

Figure 1 shows a sample text which include the answer of the query sentence.



出井伸之社長は「デジタル時代はネットワーク・ウォークマンに進化する」と未来像を語っている。「ウォークマン」の誕生は、ソニーの創業者の故井深大氏（当時・名誉会長）が海外出張の飛行機の中で聞けるカセットステレオを開発陣に頼んだのがきっかけ。特注品に注目した盛田昭夫会長（当時）が、「録音機能をはずして売れ」と商品化を指示した。ソニーが１９７９年７月に発売した携帯ヘッドホンステレオ「ウォークマン」は１日、２０周年を迎える。

Figure 1. Sample document

In this Figure, boxed parts (“出井伸之”, “井深大” and “盛田昭夫”) are person name which are recognized by NExT system and underlined parts (“社長” and “ソニー”) are clue words extracted from query sentence. In the clue words, our system recognize topical word as an important one. In the above case, “社長” can be recognized as an important one and the others are not. For each answer candidate, reciprocal number of the ranking value of word distance between the candidate and a clue word will be its score. For an important clue, the score will be double in our current implementation. Table 2 shows scores for each answer candidate with clue words. The number attached to clue words in the table means the number of appearance in the document. “ソニー 1” means the first word “ソニー” and “ソニー 2” means the second one.

Table 2. Sample score (each clue words)

	社長	ソニー 1	ソニー 2
出井伸之	2	0.5	0.33
井深大	1	1	0.5
盛田昭夫	0.67	0.33	1

And the sum of score for clue words will be the score of an answer candidate as shown in Table 3. Person name “出井伸之” will be chose for the answer of query sentence.

Table 3. Sample score (sum)

order	answer candidate	score
1	出井伸之	2.83
2	井深大	2.5
3	盛田昭夫	2

6 Results

Task1

The results were 19.4 marks out of 197.0 in TASK1 and Average Score was 0.099. Table 4 shows the output of scorer and Table 5 shows the number of correct queries in the five orders.

Table 4. Results of Task 1

Question	Answer	Output	Correct
200	497	551	43

Recall	Precision	F-measure	MRR
0.087	0.078	0.082	0.099

Table 5. The number of correct results

the order	number of correct one
first	13
second	15
third	10
forth	8
fifth	14

Task2

The results was 9.9 marks out of 197.0 in TASK2 and Average Score was 0.066. Table 6 shows the output of scorer.

Table 6. Results of Task 2

Question	Answer	Output	Correct
200	497	200	13

Recall	Precision	F-measure	MRR
0.026	0.065	0.037	0.066

7 Conclusion

In this paper, we described our QA system that consists of query pattern analysis for answer type detection and clue word extraction, document retrieval using NAMAZU system, and answer extraction based on word distance.

According to Formal Run evaluation, our QA system could not get better results but query pattern analysis and document retrieval modules are not so bad. We have roughly analyzed performance of these modules using Formal Run evaluation data. In the top five retrieved documents, almost 60% of them include correct answer. That means major problem of our QA system is answer extraction using NE system. Current NE system extracts only person name, organization name, location name and so on. In order to robust QA system, it is necessary to handle the other type of information. Improvement of answer extraction based on word distance is our future work. The other modules are also points of improvement.

References

[1] J. Fukumoto and T. Kato, An Overview of Question and Answering CHallenge (QAC) of the next NTCIR Workshop, in Proceedings of the Second NTCIR Workshop Meeting 2001.

[2] NTCIR (NII-NACSIS Test Collection for IR Systems) Project http://research.nii.ac.jp/ntcir/index-en.html

[3] Information Retrieval and Extraction Exercise (IREX) http://cs.nyu.edu/cs/projects/proteus/irex/

[4] Proceedings of 7th Message Understanding Conference (MUC-7), DARPA, 1998.

[5] ChaSen URL http://chasen.aist-nara.ac.jp/.

[6] Full-text search engine intended for easy use http://www.namazu.org/index.html.en

[7] Named Entity Extraction Tool (NExt) Homepage, http://irmscher.shiino.info.mie-u.ac.jp/next/