Question Answering System for Non-factoid Type Questions and Automatic Evaluation based on BE Method

Junichi Fukumoto
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 Japan
fukumoto@media.ritsumei.ac.jp

Abstract

In this paper, we describe answer extraction method for non-factoid questions. We classified non-factoid type questions into three types: why type, definition type and how type. We analyzed each type of questions and developed answer extraction patterns for these types of questions. For automatic evaluation, we have developed BE based evaluation tool for answers of questions. BE method is originally proposed by Hovy et. and we applied BE method for question answering evaluation. Evaluation is done by comparison between BEs of system answer and BEs of correct answers.

1 Introduction

Question Answering is a technology to find information from a huge text base using a given question. There have been evaluation workshops of question answering such as NTCIR QAC[1][Fukumoto et al. (2002b)] [Fukumoto et al. (2003)] [Fukumoto et al. (2004a)] [Kato et al. (2004b)] [Kato et al. (2004a)] [Kato et al. (2005)], TREC QA[2][Voorhees (2004)] track and CLEF[3]. In these evaluation workshops, target of given questions is mainly factoid question which requires person name, organization name, numeric expression, artifact name and so on. In TREC QA, there have applied definition type questions which require description or definition of some word or terms and other type questions which require related information of given questions.

We have already developed factoid type question answering system for QACs [Fukumoto et al. (2002a)] [Fukumoto et al. (2004b)] and also proposed question answering mechanism for why-type questions as one of non-factoid type questions [Morooka and Fukumoto (2006)]. In order to extract answers for why-type questions, we have extracted causal relations and some other relations from target documents. If one element of these relations matches question sentence, the other element will be answer for the question. We analyzed inter-sentential relations proposed in Rhetorical Structure Theory (RST) [Mann and Thompson (1987)] and have chosen causal relation, manner relation for purpose of why-type question answering. For how-type question, Nishimura et. al. proposed a method to focus on answer expressions for how-type question from Linux FAQ mailing list data [Nishimura et al. (2005)].

For QAC4, we improved why-type question answering method and expand our QA system to handle definition-type question and how-type question [Morooka and Fukumoto (2007)]. For definition-type question, we have analyzed question answer data and newspaper articles, and extracted patterns for this questions. These patterns are descriptive patterns which consist of some terms and their definition or descriptions. For how-type question, we also applied some kinds of approach as the definition-type questions. Extraction patterns for how-type questions are description of procedure.

In the evaluation of question answering, it is not so difficult to evaluate correctness of returned answer. Simple pattern matching with prepared or pooled correct answers is used for evaluation because answer string is named entity, compound noun or only noun. In the previous QACs, automatic scoring tool is used for evaluation. However, in the evaluation of non-factoid type question, evaluation will not be easy because answer string tends to be longer and has a lot of variation. In TREC2003 QA track, evaluation for definition

[1] http://www.nlp.is.ritsumei.ac.jp/qac/

[2] http://trec.nist.gov/

[3] http://clef.isti.cnr.it/

type question was done by human. Human evaluation will take a lot of cost and be difficult to keep evaluation quality in a certain level.

In text summarization, there have been several approaches to automatic evaluation of text summary. Lin proposed ROUGE [Lin (2004)] which evaluates systemsEsummaries using n-gram based statistics. In ROUGE, it is necessary to settle some parameters which suite some type of text summary and to recognize compound nouns such as named entities. Nenkova et. al. proposed Pyramid Method for evaluation of summaries in DUC [Nenkova and Passonneau (2004)]. In Pyramid Method, system summary will be broken into Summarization Content Units (SCUs) and compared them with SCUs obtained from correct summaries. However, SCU is not clearly defined and assessor sometimes provide their own SCUs. For scoring, SCU has weight according to its importance. SCU based evaluation depends on human intuition and there is ambiguity on the definition of SCUs and comparison between SCUs.

Hovy et. al. proposed an approach to automatic evaluation based on the concept of Basic Element [Hovy et al. (2006)]. Basic Element (BE) is a semantic unit (object-object relation) extracted from a sentence such as subject-object relation, modifier-object relation and so on. Evaluation of system summary using BE is based on comparison between BEs of system summary and BEs of human summary. In DUC, BE-based evaluation is utilized for evaluation of summarization and has correlation with human evaluation [Hovy et al. (2005)].

In order to apply BE-based evaluation to question answering, it is necessary to refine BE method. In QA, there are multiple answers for a given question and answer strings are various, that is, there is a case of only one noun answer or long expressions of answer. In this paper, we will describe how BE method is applied for question answering and show how BE method works in automatic evaluation of answers for questions.

2 An overview of RitsQ question answering system

We have already developed QA system for the previous QAC evaluations and we have improved our system for non-factoid question answering. For QAC4, we expanded our QA system to manage non-factoid questions, that is, expansion of question type analysis patterns for non-factoid type questions and expansion of answer extraction modules. For question type analysis, RitsQ system will analyze question type of a given question and determine its question type as why-type, definition-type or how-type according to surface expression patterns of the question if the question is non-factoid one. The surface patterns of each question type are as follows:

- Why type:

  “なぜ (naze)”, “何故 (naze)”, “どうして (doushite)”

- Definition type:

  “とは何。 (toha nani.)”, “どのような+名詞 (donoyouna + NOUN)”, “どんな+名詞 (donna + NOUN)”, “名詞+って何 (NOUN * tte nani.)”

- How type:

  “どういう (douiu)”, “どうしたら (doushitara)”, “どうする (dousuru)”, “どうすれば (dousureba)”, “どうやったら (douyattara)”, “どうやって (douyatte)”, “どうなりますか (dounarimasuka)”, “どのように (donoyouni)”, “どのような (donoyouna)”

If a given question is non-factoid and does not match to the above surface patterns, our system understand this question is definition type question as the default. For answer extraction, we prepared answer extraction module of each type of question. The details will be presented in the next section.

Another major improvement of our QA system is information retrieval module. Our previous system used Namazu system [4] for document retrieval using extracted clue words but its performance was not the level of our satisfaction. In order to improve its performance, we used information of Google snippets to re-order retrieval results of Namazu system. Firstly, we choose top 10 snippets of Google using extracted clue words. Then we calculate document similarity between retrieved top 100 documents and top 10 Google snippets to reorder the retrieved documents. We could improve IR module of our QA system because documents which include correct answers will be ranked in higher position.

[4] http://www.namazu.org/

3 Answer extraction module

Answer extraction module for non-factoid questions extracts answer strings from a paragraph of retrieved documents according to answer extraction patterns of each question type. This module searches linguistic clues of each question type for each document which is retrieved by IR module and extracts one appropriate paragraph which includes linguistic clues and some clue words of question sentence. This extracted paragraph will be a target for extraction of answer string.

3.1 Why-type question

As for why-type question, we will use the following extraction patterns and non-extraction patterns. If one sentence matches extraction patters, this sentence will be extracted as answer candidate. But this candidate will be removed from candidate list if it matches non-extraction patterns.

- extraction patterns
  - Verb + “ため (tame)”
  - Noun + “ため (tame)”
  - “ため (tame)” + Postposition “に (ni)”
  - “ため (tame)” + “、”— “。”
  - “ため (tame)” + Aux. Verb “だ (da)”
- non-extraction patterns
  - Pronoun + Postposition “の (no)” + “ため (tame)”
  - Verb + “ため (tame)” + Postposition “の (no)”
  - Noun + “ため (tame)” + Postposition “の (no)”

The semantic clue words are the words which mean reason, cause and background. This kind of words is extracted using Japanese thesaurus as follows:

“理由” “要素” “背景” “動機” “事由” “根拠” “引き金” “ゆえん” “ゆえ” “きっかけ” “悪因”

“一因” “遠因” “外因” “禍因” “画因” “起因” “基因” “近因” “偶因” “原因” “業因”

“根因” “罪因” “死因” “主因” “従因” “勝因” “心因” “真因” “成因” “善因” “素因”

“訴因” “動因” “導因” “道因” “内因” “敗因” “病因” “副因” “福因” “誘因” “要因”

We choose reason part from an extracted sentence as answer candidate. If there is conjunctive expression on the top of an extracted sentence and conjunction has a function of coordination, the previous sentence will be added in this answer candidate. If there is contradictive conjunction in a sentence, the previous part of this sentence will be removed from this answer candidate.

3.2 Definition-type question

Definition type questions require word meaning, term definition, description of term and so on. For example, in the question “What is World Heritage Convention?”, it requires definition of “World Heritage Convention” which is the most important element in this question. We call the important element Main Keyword. In order to choose Main Keyword, we firstly check blanketed word or named entity, then modifier of topic word, and finally, topic word. In the question “What agreement is World Heritage Convention?”, the word “agreement” is also important as well as Main Keyword. We call this kind of word Attributive Word. Attributive Word is the word which composes noun phrase with an interrogative such as “どういった (douitta)”, “どのような (donoyouna)”, “どんな (donna)” and so on. Extraction patterns are shown as follows:

- Main Keyword + “は (ha)” — “が (ga)” — “も (mo)”
- · · · “が (ga)” + Main Keyword + “を (wo)”
- · · · “する (suru)” + Main Keyword
- · · · “の (no)” + Main Keyword
- Main Keyword + “とは (toha)”
- · · · “の (no)” + Attributive Word
- · · · “する (suru)” + Attributive Word

If a matched sentence includes Main Keyword, the whole sentence will be an answer candidate. If a matched sentence includes Attributive Word, its modifying element will be an answer candidate.

3.3 How-type question

How-type question is inquiry of some procedure, method or conditions of action. Verbal expressions in a question sentence will be clue to recognize answer for this type of questions. For example, in the question “How is World Heritage decided?”, the verb “decide” will be important clue for answer extraction.

Extraction pattern for How-type question, we will use the main verb (Main Verb) of a question sentence and Main Keyword which is clue for Definition type question. Extraction patterns are shown as follows:

- Main Keyword + “は (ha)” — “が (ga)” + Main Verb
- “手順 (procedure)” — “手法 (method)” — “方法 (method)” — “条件 (condition)” + “は (ha)”
- “が (ga)” + “手順 (procedure)” — “手法 (method)” — “方法 (method)” — “条件 (condition)”

4 Discussion on QA system evaluation

In the evaluation, RitsQ system returned answers for 86 questions among 100 questions. For human evaluation, we select 285 answers of 86 questions by system parameter (number of system output) changing and there are 51 answers which are correct or including a part of correct answer.

As for question type analysis, there are some question sentences which our pattern failed to identify question types. For example, in the question “どういう見解を示していますか。(What kinds of opinion do you show?)”, our system recognize this question as How-type because the pattern “どういう (What kinds)” is registered in How-type, but this case should be definition type. There are several questions in the same case. It is necessary to improve question type patterns. In Why-type question, there are some errors which are caused by short of extraction patterns. We have to analyze more patterns and improve our QA system in future.

5 BE method

BE method proposed by Hovy et. al. was used for automatic evaluation of text summarization. BE is defined as a minimal semantic unit which consists of two elements and relation (head-modifier-relation) between these elements. This relation names are mainly from parse tree. In order to evaluate system summary using BE method, each sentence of system summary and reference summary will be parsed and parse tree of each summary will be broken into BEs. Evaluation is done by comparison between BEs of reference summary and BEs of system summary. If BEs of each summary are similar, system summary will be a good summary.

There are the following 4 kinds of BE Breakers provided from USC/ISI. BE Breaker is distributed as BE Package from *http://haydn.isi.edu/BE/*. In this package, BE-F system is included.

- BE-L: Chaniak parser + CYL cutting rules
- BE-F: Minipar + JF cutting rules
- Chunker: syntactic-unit chunker including cutting rules
- Microsoft parser + cutting rules

We will show an example of BE breaking using the following sentence.

Two Libyans were indicted for the Lockerbie bombing in 1991.

In this sample sentence, word “two” modifies “Libyans” and they are connected by relation “nn” (a sequence of nouns). Words “Libyans” and “indict” have relation verb-object. The results of BE breaking will be shown in Figure 1.

BE-1: (libyans, two, nn)
BE-2: (indicted, libyans, obj)
BE-3: (bombing, lockerbie, nn)
BE-4: (indicted, bombing, for)
BE-5: (bombing, 1991, in)—

Figure 1: Results of BE Breaking

There are several level of BE matching proposed by Hovy.

1. exact matching at lexical level
2. matching at the level of word original form
3. matching at the level of synonym
4. matching with paraphrase of phrase level
5. matching at semantic level

Moreover, there will be partial matching of BE elements and reference resolution of BE elements. However, current implementation of BE breaking and matching is at the level of lexical and word original form level. Hovy et al. have shown that there is correlation between evaluation by BE method and ROUGE.

6 BE-based evaluation of QA

For BE breaking of Japanese sentence, we used ChaSen for morphloigical analysis and CaboCha for syntax analysis. Figure 2 shows BE list extracted from the following sample sentence.

> ベルマーレ平塚の中田英寿が、セリエＡのペルージャへ移籍した。(Hidetoshi Nakata of Bellmare Hiratsuka moved to Perugia of Serie A.)

BE1:[中田英寿, ベルマーレ平塚, の]
BE2:[移籍した, 中田英寿, が]
BE3:[ペルージャ, セリエA, の]
BE4:[移籍した, ペルージャ, へ]

Figure 2: BE list of sample sentence

Elements of BE are independent words such as noun, verb, adjective, adverb, number and so on. Japanesen particle is used to indicate relation between elements when one element modifies the other element. If adjective modifies an element, relation between them will be modification. Table 1 summarize relations in BE.

Table 1: Results of BE Breaking

relation	meaning of relation
s	phrase with “が (ga)” — “は (ha)” modifes verb
num	numeric modifies noun or verb
mod_d	verb modifies non verb element
pro_n	pronoun modifies noun
adj	adjective modifies an element
adv	adverb modifies an element
conj	conjunction modifies an element
cae	verb modifies another verb
particle	phrase modifies an element

In case of “particle” of Table1, particle information will be relation when postpositional phrase including the particle modifies a noun.

In BE-based evaluation, system answers are scored by comparison between BEs of system answer and BEs of correct answers. Score between one system answer and one correct answer is calculated in F-measure as follows:

$$Precision(P) = \frac{matched\ BEs}{number\ of\ BEs\ of\ system}$$

$$Recall(R) = \frac{matched\ BEs}{number\ of\ BEs\ of\ correct}$$

$$F-measure = \frac{2PR}{P+R}$$

Score of one system answer will be the max score in all the scores calculated by the above F-measure for all correct answers because correct answer which has the max score will be recognized as the most similar one to the system answer. In this evaluation, if a small part of system answer is almost same as one correct answer, score of this system answer will be low. When size and contents of answers are almost the same, score will be high.

7 Experiments

In the experiment, we used RitsQ submitted results and compared BE-based evaluation with human evaluation. There are 169 answers for 64 questions among 100 given questions. In these answers, there are 18 A score, 3 B score, 11 C score and 137 D score answers. For 18 A score answers, there are 4 answers which BE evaluation score is 1 (perfect matching) but there are 6 answers which BE evaluation score 0. For 137 D score answers, there are 132 answers which BE score is 0.

We changed threshold BE score and compared human evaluation. We set threshold to 1, 0.8 and 0.6 for human A score answers as shown in Table2. All the BE score 1 answers are recognized as human score A but BE evaluation failed to recognize all the human A score answers. But if BE score level is loosed to the 0.6, BE evaluation could cover 55.6% of human A score answers (Precision 0.556) and its Recall value is still high (0.909).

We also set threshold to 0, 0.2 and 0.4 for human D score answers as shown as shown in Table3. In the all levels, the F-measure is high level (0.911 to 0.940), then BE evaluation could detect wrong answer.

Table 2: BE score A

Threshold	Precision	Recall	F-measure
1	0.222 (4/18)	1 (4/4)	0.363
0.8	0.389 (7/18)	1 (7/7)	0.560
0.6	0.556 (10/18)	0.909 (10/11)	0.690

Table 3: BE score D

Threshold	Precision	Recall	F-measure
0	0.964 (132/137)	0.917 (132/144)	0.940
0.2	0.964 (132/137)	0.863 (132/153)	0.911
0.4	1 (137/137)	0.851 (137/162)	0.920

8 Discussion on BE evaluation

In the comparison of BE evaluation and human evaluation, answers of BE evaluation score A were human score A at the threshold of more than 0.8 BE score. In answer of low BE score, the system answer includes other information which is not related to its correct answer. Such answer will have low BE score but BE scoring works well. There is a case that one system answer consists of two or more correct answers. BE score is calculated as the max value among all connect answers and then BE score of combined answer will be low. The major failure of BE scoring is exact lexical matching. Paraphrased elements will not be recognized correctly and different relation name will also not be recognized when syntax structure is different but their meanings are almost same. However, if loose matching of relation names and element names is allowed, different meanings will not be recognize. Paraphrasing of syntax structure level will be important for BE scoring.

9 Conclusion

In this paper, we describe answer extraction method for non-factoid questions. We classified non-factoid type questions into three types: why type, definition type and how type. We analyzed each type of questions and developed answer extraction patterns for these types of questions. In the evaluation of the experiment, we used question data of Formal Run of NTCIR QAC4. As a result of the experiment, our system returned 285 answers to 100 questions. We manually evaluated the result. As a result of the evaluation, our system was able to return the correct answer to 30 questions. However, performance of our QA system is not enough because of short of question patterns and answer extraction patterns for questions.

For automatic evaluation, we have developed BE based evaluation tool for answers of questions. BE method is originally proposed by Hovy et. and we applied BE method for question answering evaluation. Evaluation is done by comparison between BEs of system answer and BEs of correct answers. In the experiments, our method attained about 50% accuracy comparing in human rank A answers with F-measure scoring, and about 90% accuracy comparing in human rank D answers with F-measure scoring. In this evaluation, we have used only RitsQ system results for BE evaluation. In future, it is necessary to apply other system results submitted from the other QAC4 participants and evaluate BE method. Moreover, it is necessary to handle paraphrased elements in BE list at the level of lexical and syntax structure in order to improve performance of BE matching.

References

J. Fukumoto, T. Endo, and T. Niwa. 2002a. RitsQA: Ritsumeikan question answering system used for QAC-1. In *Working Notes of the Third NTCIR Workshop Meeting: Part IV Question Answering Challenge*, pages 113–116.

J. Fukumoto, T. Kato, and F. Masui. 2002b. Question answering challenge (QAC-1) question answering evaluation at ntcir workshop 3. In *Working Notes of the Third NTCIR Workshop Meeting: Part IV Question Answering Challenge*, pages 1–10.

J. Fukumoto, T. Kato, and F. Masui. 2003. Question Answering Challenge (QAC-1) an evaluation of question answering tasks at the NTCIR Workshop 3. In *Proc. of AAAI Spring Symposium on New Directions in Question Answering*, pages 122–133.

J. Fukumoto, T. Kato, and F. Masui. 2004a. Question answering challenge for five ranked answers and list answers - overview of NTCIR4 QAC2 Subtask 1 and 2-. In *Working Notes of the*

Fourth NTCIR Workshop Meeting, pages 283–290.

J. Fukumoto, T. Niwa, M. Itoigawa, and M. Matsuda. 2004b. Rits-QA: List answer detection and context task with zero anaphora handling. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 310–314.

E. Hovy, C.-Y. Lin, and L. Zhou. 2005. Evaluating duc 2005 using basic elements. In *Proc. of the 2005 Document Understanding Conference at NLT/EMNLP 2005*, pages –.

E. Hovy, C.-Y. Lin, L. Zhou, and J. Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proc. of the 5th International Conference on Language Resources and Evaluation*, pages –.

T. Kato, J. Fukumoto, and F. Masui. 2004a. Handling information access dialogue through qa technologies - a novel challenge for open-domain question answering. In *Proc. of the Workshop on Pragmatics of Question Answering at HLT-NAACL*, pages 70–77.

T. Kato, J. Fukumoto, and F. Masui. 2004b. Question answering challenge for information access dialogue ：- overview of NTCIR4 QAC2 Subtask 3-. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 291–297.

T. Kato, J. Fukumoto, and F. Masui. 2005. An overview of NTCIR-5 QAC3. In *Proc. of the Fifth NTCIR Workshop Meeting*, pages 361–372.

C.-Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out, Proc. of the ACL-04 Workshop*, pages 74–81.

W. C. Mann and S. A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. In *USC ISI Technical Report ISI/RS-87-190*, pages –.

K. Morooka and J. Fukumoto. 2006. Answer extraction method for why-type question answering system. In *NLC2005-107*, pages 7–12. (in Japanese).

K. Morooka and J. Fukumoto. 2007. Question answering system for non-factoid type questions. In *Proc. of the 13th Annual Meeting of ANLP*, pages 958–961. (in Japanese).

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization:the pyramid method. In *Proc. of HLT/NAACL2004*, pages –.

R. Nishimura, Y. Watanabe, and Y. Okada. 2005. A question answer system based on confirmed knowledge developed by using mails posted to a mailing list. In *IJCNLP-05*, pages –.

E. M. Voorhees. 2004. Overview of the TREC 2003 question answering track. In *Proc. of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68.