The 23rd International Technical Conference on Circuits/Systems,
Computers and Communications (ITC-CSCC 2008)

Automatic Evaluation of Question Answering System based on BE Method

Akiko Yamamoto[1] and Junichi Fukumoto[2]
[1]Graduate school of Science and Engineering, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga
525-8577 Japan
[2]Department of Media Technology, Ritsumeikan, University
1-1-1 Noji-higashi, Kusatsu, Shiga
525-8577 Japan
E-mail : [1]a_yamamoto@nlp.is.ritsumei.ac.jp, [2]fukumoto@media.ritsumei.ac.jp

Abstract: In this paper, we describe automatic evaluation method for question answering in natural language. This method is based on BEs (Basic Elements) originally proposed by Hovy et. al. for automatic evaluation of document summaries. We applied BE method for evaluation of question answering with comparison between BEs of system answer and BEs of correct answers. According to the experiments using QAC4 test set, we have proved that BE method has some correlation with human evaluation.

keywords: question answering, automatic evaluation, basic element, Pearson's correlation

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][5][2] [1], TREC[2] QA track[8][9]. In these evaluation workshops, target of given questions is mainly factoid questions which require person name, organization name, numeric expression, artifact name and so on. In the evaluation of factoid type question answering, simple pattern matching with prepared correct answers is utilized. However, in the evaluation beyond factoid type questions, such simple pattern matching cannot work because of variety of longer answer strings. In the previous evaluation workshops such as TREC QA track and NTCIR6-QAC4, evaluation was done by human and human evaluation took a lot of time and cost and was very difficult to keep evaluation quality in a certain level.

In the evaluation of document summarization, Nenkova et. al. proposed Pyramid Method for the evaluation of summaries at Document Understanding Conference (DUC)[7]. In Pyramid Method, a 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 breaking into SCUs is determined by human assessors. Then, SCU based evaluation depends on human intuition.

Hovy et. al. proposed an approach to automatic evaluation of system summaries based on Basic Element[4]. Basic Element (BE) is a basic semantic unit extracted from a sentence such as subject-object relation, modifier-object relation and so on of parse tree of target documents. In this method, system summary will be broken into a set of BEs and the evaluation will be based on comparison between BEs of a system summary and BEs of a human summary. In order to apply BE-based evaluation to question answering, it is necessary to improve 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.

[1]http://www.nlp.is.ritsumei.ac.jp/qac/
[2]http://trec.nist.gov/

2. 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.

In order to evaluate BE method, Hovy et. al. used Spearman rank correlation coefficient and Pearson's correlation and computed between DUC2003 official results[3] and BE results and showed good and consistent correlation across eval-

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

uation of different summarization tasks. 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].

3. BE-based evaluation of QA

For BE breaking of Japanese sentence, we used ChaSen [3] for morphological analysis and CaboCha [4] for syntax analysis. Then, relation names of BEs are different from English version shown in the above. 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

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.)

Elements of BE are independent words such as noun, verb, adjective, adverb, number and so on. Japanese 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.

[3]http://chasen.naist.jp/hiki/ChaSen/
[4]http://www.chasen.org/ taku/software/cabocha/

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

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:

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

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

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

In question answering, there are several ways of expressions of answer string and there are also several different answers. In BE based evaluation, we will compare all possible answers generated by human with all the system answers. One score will be calculated by the above F-measure from one human answer and system answer. If there are M kinds of system answer and N kinds of human answers, all the possible combination (M $\times$ N) of system answer and human answer will calculated and the best score will be the score of the question. 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.

3.1 Loose pattern matching

Basically, BE matching is done by exact matching of BE elements (two elements and relation between them), that is, element and relation expressions are literally the same ones. However, there are some cases that literal expressions of BE elements are different but these elements are semantically equal. For example, a noun phrase is repeated in different forms of a document. Verb phrases which have different tense forms could be recognized as the semantically same one. When a relation name of BE is Japanese particle between elements, semantically similar particles in different surface expressions could be also recognized as the same one.

We have introduced loose pattern matching into BE matching of BE-based evaluation. Loose pattern matching will be done according to types of BE elements: noun phrase type and verb phrase type. In the both types of elements, when core words in the elements are shared, these elements could be recognized as matched ones in loose pattern matching. Loose pattern matching method of both types of elements is shown as follows:

- noun phrase type
  If the last words of noun phrase elements are the same one, these elements of this type will be loose pattern matched. This is because a head element is frequently located in the last position of Japanese noun phrase.
- verb phrase type
  If main verbs or adjectives of verb phrasal elements are the same ones, these elements will be loose pattern matched. In Japanese verb phrase, the main element of a verbal phrase is verb or adjectives and auxiliaries are not the main element.

For example, BE elements "与えた (atae-ta) gave" and "与える (atae-ru) give" will be matched because the main verbs of these elements are same and only their infection forms are different. If this verb phrase includes auxiliary verb "た (ta)" which means past tense, this verb phrase will be matched with the verb phrase "与える (atae-ru) give" which means present tense. This is because tense information is not related their contents level information. As for noun phrase type, BE element "市民団体 (shimin danntai) a community group" and "団体 (danntai) a group" will be matched because their head nouns "団体 (danntai) a group" are the same one.

4. Experiments

4.1 Data for Experiments

In the experiment, we used question answering data which was used for QAC4 Formal Run[2]. There were 100 questions which required longer answer expressions such as answers for why-type question and so on. These answers were extracted from the target documents which were two years (1998 and 1999) Mainichi newspaper articles. For these 100 questions, 1071 correct answer were prepared for QAC4 evaluation. There were 16 systems from all the task participants and 3750 system answer for the 100 questions were sent from these participants.

For evaluation of these system answers, two human assessors have evaluated the system answers based on the prepared correct answers according to the following evaluation criterion.

- Level A:
  System answer has almost the same contents as one of the correct answers and there is few information expect for the same contents in the answer. If there is an additional expression which has no effect on the contents, this case is recognized in this level.
- Level B:
  System answer includes the contents of one of the correct answers and the other information, and the main contents are not the contents of the correct answer.
- Level C:
  System answer includes some part (not all one) of the contents of the correct answers.
- Level D:
  System answer includes no information of any of the contents of the correct answers. There is a case that some surface expression of the correct answer is included in the system answer. If this expression is used for the other meaning, this case will be Level D. If this expression is used for the same meaning of a part of the correct answer, this case will be Level C.

4.2 Loose pattern matching

Firstly, we evaluate the efficiency of loose pattern matching in the experiments. We broken all the system answers and correct answers into both BE sets. Then, we compared BE set of system answers and BE set of correct answers with loose pattern patching. 442 BE pairs are matched in the loose patterns. Among the matched 442 BE pairs, we checked adequacy of these matching by human and 430 pairs are semantically correct as the results. Therefore, our loose pattern matching attained almost 97% accuracy in BE matching.

4.3 BE-base evaluation

We have evaluated how BE-base automatic evaluation system simulates human evaluation. For this evaluation, we used Pearson's correlation between systems' results and correct answers using F-measure of BE matching values. BE matching value is calculated by matching ratio between BEs of a system answer and BEs of a correct answer. Precision value will be obtained from an average of BE matching values of all the system answers. BE matching value of an system answer will be the highest values of BE matching values for all correct answers. Recall value will be obtained from an average of BE matching values of all the correct answers. The BE matching value of a correct answer will be the highest value of BE matching values for all system answers. The following formulas show Precision (P_{BE}), Recall (R_{BE}) and $F-measure_{BE}$ of BE results.

$$P_{BE} = \frac{sum\ of\ BE\ values\ of\ system\ answers}{number\ of\ system\ answers}$$

$$R_{BE} = \frac{sum\ of\ BE\ values\ of\ correct\ answers}{number\ of\ correct\ answers}$$

$$F - measure_{BE} = \frac{2 * P_{BE} * R_{BE}}{P_{BE} + R_{BE}}$$

In order to calculate human evaluation values given by human assessors, we set values 1.0, 0.5, 0.5 and 0.0 for Level A, B, C and D, respectively. Human evaluation values will be calculated by replacing BE matching values and human evaluation values according to the above formulas. Precision (P_{hum}), Recall (R_{hum}) and $F-measure_{hum}$ of human evaluation results are obtained in the following formulas.

$$P_{hum} = \frac{sum\ of\ human\ values\ of\ system\ answers}{number\ of\ system\ answers}$$

$$R_{hum} = \frac{sum\ of\ human\ values\ of\ correct\ answers}{number\ of\ correct\ answers}$$

$$F - measure_{hum} = \frac{2 * P_{hum} * R_{hum}}{P_{hum} + R_{hum}}$$

Table 2. Pearson score between system and human

pattern matching	Pearson correlation score
exact pattern matching	0.784
loose pattern matching	0.813

For each question data, we have calculated Pearson's correlation between systems' results and correct answers using the above F-measure values as shown in Table2.

When we use ordinal exact pattern matching of BEs, Pearson's correlation almost attained 0.8. Moreover, in case of loose pattern matching, Pearson's correlation exceeds 0.8. We can say that automatic evaluation using BE method works well and BE method can be used for QA evaluation instead of human evaluation. The correlation values of loose pattern matching was higher than exact pattern matching, therefore, BE method using loose pattern matching is more like human evaluation of QA.

5. Discussions

5.1 BE-based evaluation

In the comparison between BE evaluation and human evaluation using Pearson's correlation, BE-based evaluation of question answering works well using both exact pattern matching and loose pattern matching. 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 correct answers and then BE score of combined answer will be low.

5.2 Loose pattern matching

We have introduced loose pattern matching to resolve some problems in exact lexical matching. According to evaluation of loose pattern matching, it covers some lexical variation of pattern matching without any problems. However, there are still some problems in lexical pattern matching of BE scoring. 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.

We will show some failure examples in loose pattern matching. When a verb phrase includes negation expression which is expressed in Japanese auxiliary verb, such verb phrase will be matched with a verb phrase without negation. In this case, loose pattern could not work. In our current implementation, head verb is only handled in pattern matching, then opposite meaning elements are recognized as the same one. In noun phrase type, when a BE element includes paraphrased nouns, BE matching will fail. For example, "2008 年 (2008 nen) 2008 year" is expressed in abbreviation form such as "08 年 (08 nen) 2008 year". It is necessary to expand loose pattern matching rules to paraphrasing.

6. Conclusion

In this paper, we applied BE method for question answering evaluation and have developed a method to compare system answers and correct answers for evaluation. Evaluation was done by comparison between BEs of a system answer and BEs of a correct answers. In the experiments using QAC4 testset, we have proved that BE method has correlation with human evaluation based on Pearson ' s correlation. We have improved pattern matching of BE elements in our BE evaluation method. Evaluation using loose pattern matching got better results than exact matching. Then, BE-based evaluation of question answering works well and we can use BE-based evaluation instead of human evaluation. However, there are some cases that loose pattern matching could not work. 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.

Acknowledgements

We would like to thank all the participants to the NTCIR-6 QAC task. We also express my thanks to QAC task organizers and NTCIR organizers.

References

[1] J. Fukumoto, T. Kato, and F. Masui. 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, 2004.

[2] J. Fukumoto, T. Kato, F. Masui, and T. Mori. An overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6. In *Proc. of the Sixth NTCIR Workshop Meeting*, pages 433–440, 2007.

[3] E. Hovy, C.-Y. Lin, and L. Zhou. Evaluating DUC 2005 using Basic Elements. In *Proc. of the 2005 Document Understanding Conference at NLT/EMNLP 2005*, 2005.

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

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

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

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

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

[9] E. M. Voorhees. Overview of the TREC 2004 question answering track. In *Proc. of the Thirteenth Text REtrieval Conference (TREC 2004)*, pages 53–62, 2005.