Answer Validation by Keyword Association

Masatsugu Tonoike, Takehito Utsuro and Satoshi Sato
Graduate school of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku 606-8501 Kyoto, JAPAN
{tonoike,utsuro,sato}@pine.kuee.kyoto-u.ac.jp

Abstract

Answer validation is a component of question answering system, which selects reliable answer from answer candidates extracted by certain methods. In this paper, we propose an approach of answer validation based on the strengths of lexical association between the keywords extracted from a question sentence and each answer candidate. The proposed answer validation process is decomposed into two steps: the first is to extract appropriate keywords from a question sentence using word features and the strength of lexical association, while the second is to estimate the strength of the association between the keywords and an answer candidate based on the hits of search engines. In the result of experimental evaluation, we show that a good proportion (79%) of a multiple-choice quiz "Who wants to be a millionaire" can be solved by the proposed method.

1 Introduction

The technology of searching for the answer of a question written in natural language is called "Question Answering"(QA), and has gotten a lot of attention recently. Research activities of QA have been promoted through competitions such as TREC QA Track (Voorhees, 2004) and NTCIR QAC (Fukumoto et al., 2004). Question answering systems can be decomposed into two steps: first step is to collect answer candidates, while the second is to validate each of those candidates. The first step of collecting answer candidates has been well studied so far. Its standard technology is as follows: first, the answer type of a question, such as LOCATION or PERSON, is identified. Then, the documents which may contain answer candidates are retrieved by querying available document set with queries generated from the question sentence. Finally, named entities which match the answer type of the question sentence are collected from the retrieved documents as answer candidates.

In this paper, we focus on the second step of how to validate an answer candidate. Several answer validation methods have been proposed. One of the well-known approaches is that based on deep understanding of text (e.g. Moldovan et al. (2003)). In the approach of answer validation based on deep understanding, first a question and the paragraph including an answer candidate are parsed and transformed into logical forms. Second, the validity of the answer candidate is examined through logical inference. One drawback of this approach is that it requires a rich set of lexical knowledge such as WordNet and world knowledge such as the inference rule set. Consequently, this approach is computationally expensive. In contrast, in this paper, we propose another approach of answer validation, which is purely based on the estimation of the strengths of lexical association between the keywords extracted from a question sentence and each answer candidate. One underlying motivation of this paper is to examine the effectiveness of quite low level semantic operation such as measuring lexical association against knowledge rich NLP tasks such as answer validation of question answering. Surprisingly, as we show later, given multiple-choices as answer candidates of a question, a good proportion of a certain set of questions can be solved by our method based on lexical association.

In our framework of answer validation by keyword association (in the remaining of this paper, we call the notion of the lexical association introduced above as "*keyword association*"), the answer validation process is decomposed into two steps: the first step is to extract appropriate keywords from a question sentence, while the second step is to estimate the strength of the association between the keywords and an answer candidate. We propose two methods for the keyword selection step: one is by a small number of hand-crafted rules for determining word weights based on word features, while the other is based on search engine hits. In the second step of how to validate an answer candidate, the web is used as a knowledge base for estimating the strength of the association between the extracted keywords and an answer candidate.

Its basic idea is as follows: the stronger the association between the keywords and an answer candidate, the more frequently they co-occur on the web. In this paper, we introduce several measures for estimating the strength of the association, and show their effectiveness through experimental evaluation.

In this paper, in order to concentrate on the issue of answer validation, but not the whole QA processes, we use an existing multiple-choice quiz as the material for our study. The multiple-choice quiz we used is taken from "Who wants to be a millionaire". "Who wants to be a millionaire" is a famous TV show, which originated in the United Kingdom and has been localized in more than fifty countries. We used the Japanese version, which is produced by Fuji Television Network, Inc.. In the experimental evaluation, about 80% of the questions of this quiz can be solved by the proposed method of answer validation by keyword association.

Section 2 introduces the idea of question answering by keyword association. Section 3 describes how to select keywords from a question sentence. Section 4 describes how to select the answer of multiple-choice questions. Section 5 describes how to integrate the procedures of keyword selection and answer selection. Section 6 presents the results of experimental evaluations. Section 7 compares our work with several related works Section 8 presents our conclusion and future works.

2 Answer Validation by Keyword Association

2.1 Keyword Association

Here is an example of the multiple-choice quiz.

Q1: Who is the director of "American Graffiti"?

a: George Lucas
b: Steven Spielberg
c: Francis Ford Coppola
d: Akira Kurosawa

Suppose that you do not know the correct answer and try to find it using a search engine on the Web. The simplest way is to input the query "American Graffiti" to the search engine and skim the retrieved pages. This strategy assumes that the correct answer may appear on the page that includes the keyword "American Graffiti". A little cleverer way is to consider the number of pages that contain both the keyword and a choice. This number can be estimated from the hits of a search engine when you input a conjunct query "American Graffiti" and "George Lucas". Based on this assumption, it is reasonable to hypothesize that the choice which has the largest hits is the answer. For the above question Q1, this strategy works. Table 1 shows the hits of the conjunct queries for each of the choices. We used "google[1]" as a search engine. Here, let X be the set of keywords, Y be the choice. Function *hits* is defined as follows.

Table 1: Hits of Keywords and the Choices for the Question Q1 (X:"American Graffiti")

Y(choice)	hits(X and Y)
"George Lucas"	15,500
"Steven Spielberg"	5,220
"Francis Ford Coppola"	4,800
"Akira Kurosawa"	836

$$hits(X) \equiv hits(x_1 \text{ AND } x_2 \text{ AND} \cdots \text{AND } x_n)$$

where

$$X = \{x_1, x_2, \ldots, x_n\}$$

The conjunct query with "George Lucas", which is the correct answer, returns the largest hits.

Here, the question Q1 can be regarded as a question on the strength of association between keyword and an choice, and converted into the following form.

Q1′: Select the one that has the strongest association with "American Graffiti".

a: George Lucas
b: Steven Spielberg
c: Francis Ford Coppola
d: Akira Kurosawa

We call this association between the keyword and the choice as *keyword association*.

2.2 How to Select Keywords

It is important to select appropriate keywords from a question sentence. Consider the following question.

Q2: Who is the original author of the famous movie "Lord of the Rings"?

a: Elijah Wood
b: JRR Tolkien
c: Peter Jackson
d: Liv Tyler

The numbers of hits are shown in Table 2. Here, let X be "Lord of the Rings", X' be "Lord of the

[1]http://www.google.com

Table 2: Hits of Keywords and the Choices for the Question Q2 (X:"Lord of the Rings", X':"Lord of the Rings" and "original author")

Y(choice)	hits (X and Y)	hits (X' and Y)
"Elijah Wood"	682,000	213
"JRR Tolkien"	652,000	702
"Peter Jackson"	1,140,000	340
"Liv Tyler"	545,000	106

Rings" and "original author". When you select the title of this movie "Lord of the Rings" as a keyword, the choice with the maximum hits is "Peter Jackson", which is not the correct answer "JRR Tolkien". However, if you select "Lord of the Rings" and "original author" as keywords, this question can be solved by selecting the choice with maximum hits. Therefore, it is clear from this example that how to select appropriate keywords is important.

2.3 Forward and Backward Association

For certain questions, it is not enough to generate a conjunct query consisting of some keywords and a choice, and then to simply select the choice with maximum hits. This section introduces more sophisticated measures for selecting an appropriate answer. Consider the following question.

Q3: Where is Pyramid?

a: Canada
b: Egypt
c: Japan
d: China

The numbers of hits are shown in Table 3. In this case, given a conjunct query consisting of a keyword "Pyramid" and a choice, the choice with the maximum hits, i.e., "Canada" is not the correct answer "Egypt". Why could not this question be solved? Let us consider the hits of the choices alone. The hits of the atomic query "Canada" is about seven times larger than the hits of the atomic query "Egypt". With this observation, we can hypothesize that the hits of a conjunct query "Pyramid" and a choice are affected by the hits of the choice alone. Therefore some normalization might be required.

Based on the analysis above, we employ the metrics proposed by Sato and Sasaki (2003). Sato and Sasaki (2003) has proposed two metrics for evaluating the strength of the relation of two terms. Suppose that X be the set of keywords and Y be the choice. In this paper, we call the hits of a conjunct query consisting of keywords X and a choice Y, which is normalized by the hits of X, as *forward association* $FA(X,Y)$. We also call the hits of a conjunct query X and Y, which is normalized by the hits of Y, as *backward association* $BA(X,Y)$.

$$FA(X,Y) = hits(X \cup \{Y\})/hits(X)$$
$$BA(X,Y) = hits(X \cup \{Y\})/hits(\{Y\})$$

Note that when X is fixed, $FA(X,Y)$ is proportional to $hits(X \cup \{Y\})$.

Let's go back to Q3. In this case, the choice with the maximum BA is correct. Some questions may solved by referring to FA, while others may be solved only by referring to BA. Therefore, it is inevitable to invent a mechanism which switches between FA and BA.

2.4 Summary

Based on the observation of Sections 2.1 $\sim$ 2.3, the following three questions must be addressed by answer validation based on keyword association.

- How to select appropriate keywords from a question sentence.
- How to identify the correct answer considering forward and/or backward association.
- How many questions can be solved by this strategy based on keyword association.

3 Keyword Selection

This section describes two methods for selecting appropriate keywords from a question sentence: one is based on the features of each word, the other based on hits of a search engine.

First, all the nouns are extracted from the question sentence using a Japanese morphological analyzer JUMAN(Kurohashi and Nagao, 1999) and a Japanese parser KNP(Kurohashi, 1998). Here, when the sequence of nouns constitute a compound, only the longest compound is extracted and their constituent nouns are not extracted. Let N denote the set of those extracted nouns and compounds, from which keywords are selected. In the following, the search engine "goo[2]" is used for obtaining the number of hits.

[2]http://www.goo.ne.jp

Table 3: Hits of Keywords and the Choices for the Question Q3

X(keyword)	hits(X)			
Pyramid	3,170,000			
Y(choice)	hits(Y)	hits(Y and X)	FA(X, Y)	BA(X, Y)
Canada	100,000,000	334,000	0.105	0.00334
Egypt	14,500,000	325,000	0.103	0.0224
Japan	63,100,000	246,000	0.0776	0.00390
China	53,600,000	225,000	0.0710	0.00420

3.1 Keyword Selection Based on Word Features

In this method, keywords are selected by the following procedure:

1. If the question sentence contains n quotations with quotation marks '「' and '」', those n quoted strings are selected as keywords.
2. Otherwise:
   2-1. According to the rules for word weights in Table 4, weights are assigned to each element of the keyword candidate set N.
   2-2. Select the keyword candidate with the maximum weight and that with the second maximum weight.
   2-3. i. If the hits of AND search of those two keyword candidates are 15 or more, both are selected as keywords.
   ii. Otherwise, select the one with the maximum weight.

Let k denote the set of the selected keywords ($k \subseteq N$), we examine the correctness of k as follows. Let c denote a choice, $c_1^{FA}(k)$ the choice with the maximum $FA(k, c)$, and $c_1^{BA}(k)$ the choice with the maximum $BA(k, c)$, respectively.

$$c_1^{FA}(k) = \mathop{\mathbf{argmax}}_{c} FA(k, c)$$
$$c_1^{BA}(k) = \mathop{\mathbf{argmax}}_{c} BA(k, c)$$

Here, we regard the selected keywords k to be *correct* if either $c_1^{FA}(k)$ or $c_1^{BA}(k)$ is correct. Against the development set which is to be introduced in Section 6.1, the correct rate of the keywords selected by the procedure above is 84.5%.

Table 4: Rules for Word Weights

rule	weight
n-th segment	$(1 + 0.01 \times n)$
stopword	0
quoted by quotation marks 「」	3
person name	3
verbal nouns ('sahen'-verb stem)	0.5
word which expresses relation	2
Katakana	2
name of an award	2
name of an era	0.5
name of a country	0.5
number	3
hits > 1000000	
and consists of one character	0.9
marked by a topic maker and	
name of a job	0.1
hits > 100000	0.2
hits < 10000	1.1
number of characters = 1	0.2
number of characters = 2	0.25
number of characters = 3	0.5
number of characters = 4	1.1
number of characters ≥ 5	1.2

3.2 Keyword Selection Based on Hits of Search Engine

3.2.1 Basic Methods

First, we introduce several basic methods for selecting keywords based on hits of a search engine. Let 2^N denote the power set of N, where a set of keywords k is an element of 2^N ($k \in 2^N$). Let $\hat{k}$ denote the selected set of keywords and $\hat{c}$ the selected choice.

The first method is to simply select the pair of $\langle \hat{k}, \hat{c} \rangle$ which gives the maximum hits as below:

$$\langle \hat{k}, \hat{c} \rangle = \mathop{\mathbf{argmax}}_{c,\ k \in 2^N} hits(k \cup \{c\})$$

Against the development set, the correct rate of the choice which is selected by this method is 35.7%.

In a similar way, another method which se-

lects the maximum FA or BA can be given as below:

$$\langle \hat{k}, \hat{c} \rangle = \underset{c,\ k \in 2^N}{\mathbf{argmax}}\, FA(k \cup \{c\})$$

$$\langle \hat{k}, \hat{c} \rangle = \underset{c,\ k \in 2^N}{\mathbf{argmax}}\, BA(k \cup \{c\})$$

Their correct rates are 71.3% and 36.1%, respectively.

3.2.2 Keyword Association Ratio

Next, we introduce more sophisticated methods which use the ratio of maximum and second maximum associations such as FA or BA. The underlying assumption of those methods are that: the greater those ratios are, the more reliable is the selected choice with the maximum FA/BA. First, we introduce two methods: FA ratio and BA ratio.

FA ratio This is the ratio of FA of the choice with second maximum FA over one with maximum FA. FA ratio is calculated by the following procedure.

1. Select the choices with maximum FA and second maximum FA.
2. Estimate the correctness of the choice with maximum FA by the ratio of their FAs.

The set $\hat{k}$ of keywords and the choice $\hat{c}$ to be selected by FA ratio are expressed as below:

$$\hat{k} = \underset{k \in 2^N}{\mathbf{argmin}} \frac{FA(k, c_2^{FA}(k))}{FA(k, c_1^{FA}(k))}$$

$$\hat{c} = c_1^{FA}(\hat{k})$$

$$c_2^{FA}(k) = \underset{c}{\mathbf{arg\text{-}secondmax}}\, FA(k, c)$$

where $\mathbf{arg\text{-}secondmax}_c$ is defined as a function which selects c with second maximum value.

Similarly, the method based on BA ratio is given as below:

BA ratio

$$\hat{k} = \underset{k \in 2^N}{\mathbf{argmin}} \frac{BA(k, c_2^{BA}(k))}{BA(k, c_1^{BA}(k))}$$

$$\hat{c} = c_1^{BA}(\hat{k})$$

$$c_2^{BA}(k) = \underset{c}{\mathbf{arg\text{-}secondmax}}\, BA(k, c)$$

Unlike the methods based on FA ratio and BA ratio, the following two methods consider both FA and BA. The motivation of those two methods is to regard the decision by FA and BA to be reliable if FA and BA agree on selecting the choice.

Table 5: Evaluation of Keyword Association Ratios (precision/coverage)(%)

		max and second max	
		FA	BA
ratio	FA	63.1/100	70.6/95.0
	BA	**75.8/93.2**	67.6/100

BA ratio with maximum and second maximum FA

$$\hat{k} = \underset{k \in 2^N}{\mathbf{argmin}} \frac{BA(k, c_2^{FA}(k))}{BA(k, c_1^{FA}(k))}$$

$$\hat{c} = c_1^{FA}(\hat{k})$$

FA ratio with maximum and second maximum BA

$$\hat{k} = \underset{k \in 2^N}{\mathbf{argmin}} \frac{FA(k, c_2^{BA}(k))}{FA(k, c_1^{BA}(k))}$$

$$\hat{c} = c_1^{BA}(\hat{k})$$

Coverages and precisions of these four methods against the development set are shown in Table 5. Coverage is measured as the rate of questions for which the ratio is less than or equal to 1[3]. Precisions are measured as the rate of questions for which the selected choice $\hat{c}$ is the correct answer, over those covered questions. The method having the greatest precision is BA ratio with maximum and second maximum FA. In the following sections, we use this ratio as the keyword association ratio. Table 6 farther examines the correlation of the range of the ratio and the coverage/precision. When the ratio is less than or equal to 0.25, about 60% of the questions are solved with the precision close to 90%. This threshold of 0.25 is used in the Section 5 when integrating the keyword association ratio and word weights.

4 Answer Selection

In this section, we explain a method to identify the correct answer considering forward and/or backward association. After selecting keywords, the following numbers are obtained by a search engine.

- Hits of the keywords X: $hits(X)$
- Hits of the choice Y: $hits(\{Y\})$

[3]For the ratios considering both FA and BA, the ratio greater than 1 means that FA and BA disagree on selecting the choice.

Table 6: Evaluation of Keyword Association Ratio: BA ratio of FA max and second-max

ratio	# of questions	
	coverage	precision
0	18.9% (163/888)	89.6% (146/163)
≤ 0.01	21.5% (191/888)	89.5% (171/191)
≤ 0.1	40.5% (360/888)	87.5% (315/360)
$\mathbf{\leq 0.25}$	**60.4% (536/888)**	**86.9% (466/536)**
≤ 0.5	78.0% (693/888)	81.6% (566/693)
≤ 0.75	87.2% (774/888)	78.4% (607/774)
≤ 1	93.2% (828/888)	75.8% (628/828)

- Hits of the conjunct query: $hits(X \cup \{Y\})$

Then for each choice Y, FA and BA are calculated. As introduced in section 3, $c_1^{FA}(k)$ denotes the choice whose FA value is highest, and $c_1^{BA}(k)$ the choice whose BA value is highest. What has to done here is to decide which of $c_1^{FA}(k)$ and $c_1^{BA}(k)$ is correct.

After manually analyzing the search engine hits against the development set, we handcrafted the following rules for switching between $c_1^{FA}(k)$ and $c_1^{BA}(k)$.

1. if $c_1^{FA}(k) = c_1^{BA}(k)$ then $c_1^{FA}(k)$
2. else if $\frac{FA(k,c_1^{BA}(k))}{FA(k,c_1^{FA}(k))} \geq 0.8$ then $c_1^{BA}(k)$
3. else if $\frac{FA(k,c_1^{BA}(k))}{FA(k,c_1^{FA}(k))} \leq 0.2$ then $c_1^{FA}(k)$
4. else if $\frac{BA(k,c_1^{FA}(k))}{BA(k,c_1^{BA}(k))} \geq 0.53$ then $c_1^{FA}(k)$
5. else if $hits(k) \geq 1300$ then $c_1^{BA}(k)$
6. else if $\frac{FA(k,c_1^{BA}(k))}{FA(k,c_1^{FA}(k))} \geq 0.6$ then $c_1^{BA}(k)$
7. else $c_1^{FA}(k)$

Table 7 shows the results of evaluating precision of answer selection methods against the development set, when the keywords are selected based on word weights in Section 3.1. In the table, in addition to the result of answer selection rules above, the results with baselines of selecting the choice with maximum FA or BA are also shown. It is clear from the table that the answer selection rules described here significantly outperforms those baselines.

For each of the answer selection rules, Table 8 shows its precision. In the development set[4], there are 541 questions (about 60%) where

[4]Four questions are excluded because hits of the conjunct query $hits(X \cup \{Y\})$ were 0

Table 7: Precision of Answer Selection (with keyword selection by word weights)

method	precision
max FA	70.8%
max BA	67.6%
selection rule	77.3%

Table 8: Evaluation of Each Answer Selection Rule (with keyword selection by word weights)

rule	answer	precision
1	$c_1^{FA}(k) = c_1^{BA}(k)$	88.5% (479/541)
$2 \sim 6$	-	60.3% (207/343)
total	-	77.6% (686/884)
2	$c_1^{BA}(k)$	65.3% (32/49)
3	$c_1^{FA}(k)$	61.8% (68/110)
4	$c_1^{FA}(k)$	53.6% (37/69)
5	$c_1^{BA}(k)$	60.3% (35/58)
6	$c_1^{BA}(k)$	66.7% (12/18)
7	$c_1^{FA}(k)$	59.0% (23/39)

$c_1^{FA}(k)$ and $c_1^{BA}(k)$ are identical, and the 88.5% of the selected choices are correct. This result shows that more than half of the questions $c_1^{FA}(k)$ is equal to $c_1^{BA}(k)$ and about 90% of these questions can be solved. This result shows that whether FA and BA agree or not is very important and is crucial for reliably selecting the answer.

5 Total Procedure of Keyword Selection and Answer Selection

Finally, the procedures of keyword selection and answer selection presented in the previous sections are integrated as given below:

1. If $ratio \leq 0.25$:
   Use the set of keywords selected by BA ratio with maximum and second maximum FA. The choice to be selected is the one with maximum BA.
2. Otherwise:
   Use the set of keywords selected by word weights. Answer selection is done by the procedure of Section4.

6 Evaluation

6.1 Data Set

In this research, we used the card game version of "クイズ$ミリオネア (Who wants to be a millionaire)", which is sold by Tomy Company,

LTD. It has 1960 questions, which are classified into fifteen levels according to the amount of prize money. Each question has four choices. All questions are written in Japanese. The followings give a few examples.

10,000 yen level
[A39] エジプトやケニアなどの国がある大陸はどこ？
(Which continent are Egypt and Kenya located in?)
A. アフリカ大陸 (Africa)
B. ユーラシア大陸 (Eurasia)
C. 北アメリカ大陸 (North America)
D. 南アメリカ大陸 (South America)
[Correct Answer: アフリカ大陸]

1,000,000 yen level
[J39] コロンブスが新大陸発見の航海をしたときに乗っていた船の名前は何？
(What is the name of the ship in which Columbus was sailing when he discovered a new continent?)
A. アトランティス号 (Atlantis)
B. アルゴ号 (Argo)
C. サンタマリア号 (Santa Maria)
D. ノーチラス号 (Nautilus)
[Correct Answer: サンタマリア号]

10,000,000 yen level
[O4] 夏期オリンピック大会で参加国数が初めて１００ヶ国を越えた大会はどれ？
(In which summer Olympics did the number of participating countries first exceed 100?)
A. ローマ五輪 (Rome Olympics)
B. 東京五輪 (Tokyo Olympics)
C. メキシコ五輪 (Mexico Olympics)
D. ミュンヘン五輪 (Munich Olympics)
[Correct Answer: メキシコ五輪]

We divide questions of each level into two halves: first of which is used as the development set and the second as the test set. We exclude questions with superlative expressions (e.g., Out of the following four countries, select the one with the maximum number of states.) or negation (e.g., Out of the following four colors, which is not used in the national flag of France.) because they are not suitable for solving by keyword association. Consequently, the development set comprises 888 questions, while the test set comprises 906 questions. The number of questions per prize money amount is shown in Table 9.

Table 9: The number of questions per prize money amount

prize money amount (yen)	# of questions		
	full	dev	test
10,000	160	71	74
20,000	160	71	77
30,000	160	67	70
50,000	160	75	71
100,000	160	73	73
150,000	160	76	72
250,000	160	71	77
500,000	160	74	77
750,000	160	78	71
1,000,000	160	73	76
1,500,000	120	53	58
2,500,000	90	38	42
5,000,000	70	30	32
7,500,000	50	24	21
10,000,000	30	14	15
total	1960	888	906

We compare the questions of "Who wants to be a millionaire" with those of TREC 2003 QA track and those of NTCIR4 QAC2 task. The questions of "Who wants to be a millionaire" are all classified as factoid question. They correspond to TREC 2003 QA track factoid component. The questions of NTCIR4 QAC2 are also all classified as factoid question. We compare bunsetsu [5] count of the questions of "Who wants to be a millionaire" with word count of the questions of TREC 2003 QA track factoid component and bunsetsu count of the questions of NTCIR4 QAC2 Subtask1. The questions of "Who wants to be a millionaire" consist of 7.24 bunsetsu on average, while those of TREC 2003 QA track factoid component consist of 7.76 words on average, and those of NTCIR4 QAC2 Subtask1 consist of 6.19 bunsetsu on average. Therefore, it can be concluded that the questions of "Who wants to be a millionaire" are not shorter than those of TREC 2003 QA track and those of NTCIR4 QAC2 task.

6.2 Results

Against the development and the test sets, Table 10 shows the results of evaluating the total procedure of keyword selection and answer selection presented in Section 5. The table also shows the performance of baselines:

[5] A bunsetsu is one of the linguistic units in Japanese. A bunsetsu consists of one content word possibly followed by one or more function words.

Table 10: Total Evaluation Results (precision/coverage)(%)

method	dev	test
K.A.R. ($r \leq 1$)	75.8/93.2	74.6/93.6
word weights + answer selection	77.3/100	73.4/100
Integration	78.6/100	75.9/100
K.A.R. ($r \leq 0.25$)	86.9/60.4	86.0/61.5
word weights ($r > 0.25$) + answer selection	65.9/39.6	59.9/38.5

K.A.R.: keyword association ratio

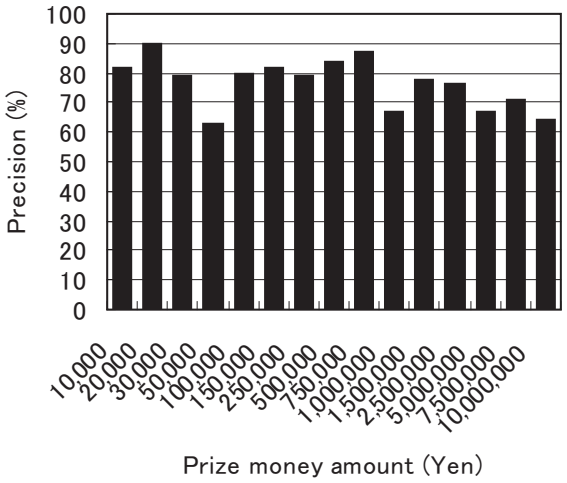

Figure 1: Precision classified by prize money amount

i.e., keyword association ratio presented in Section 3.2.2, and word weights of Section 3.1 + answer selection of Section 4. Integration of keyword association ratio and word weight outperforms those baselines. In total, about 79% (for the development set) and 76% (for the test set) of the questions are solved by the proposed answer validation method based on keyword association.

Comparing the performance of the two data sets, word weights + answer selection has 4% lower precision in the test set. This result indicates that rules for word weights as well as answer selection rules overfit to the development set. On the other hand, the difference of the precisions of the keyword association ratio is much less between the two data sets, indicating that keyword association ratio has less overfit to the development set.

Finally, the result of the experiment where the development set was solved by the integration method was classified by prize money amount. The result is shown in Figure 1. The more the prize money amount is, the lower the precision seems to be, while their precisions are all above 60%, and their differences are less than 20% in most cases. It can be concluded that our system can solve questions of all the levels almost equally.

7 Related Work

Kwok et al. (2001) proposed the first automated question-answering system which uses the web. First, it collects documents that are related to the question sentence using google and picks answer candidates up from them. Second, it selects an answer based on the frequency of candidates which appear near the keywords.

In the method proposed by Brill et al. (2002), answer candidates are picked up from the summary pages returned by a search engine. Then, each answer candidate is validated by searching for relevant documents in the TREC QA document collection. Both methods do not consider the number of hits returned by the search engine.

Magnini et al. (2002) proposed an answer validation method which uses the number of search engine hits. They formulate search engine queries using AltaVista's OR and NEAR operators. Major difference between the method of Magnini et al. (2002) and ours is in keyword selection. In the method of Magnini et al. (2002), the initial keywords are content words extracted from a question sentence. If the hits of keywords is less than a threshold, the least important keyword is removed. This procedure is repeated until the hits of the keywords is over the threshold. On the other hand, in our method, keywords are selected so that the strength of the association between the keyword and an answer candidate is maximized. Intuitively, our method of keyword selection is more natural than that of Magnini et al. (2002), since it considers both the question sentence and an answer candidate. As for measures for scoring answer candidates, Magnini et al. (2002) proposed three measures, out of which "Corrected Conditional Probability" performs best. In our implementation, the performance of "Corrected Conditional Probability" is about 5% lower than our best result.

8 Conclusion and Future Work

In this paper, we proposed an approach of answer validation based on the strengths of lexical association between the keywords extracted

from a question sentence and each answer candidate. The proposed answer validation process is decomposed into two steps: the first is to extract appropriate keywords from a question sentence using word features and the strength of lexical association, while the second is to estimate the strength of the association between the keywords and an answer candidate based on the hits of search engines. In the result of experimental evaluation, we showed that a good proportion (79%) of the multiple-choice quiz "Who wants to be a millionaire" can be solved by the proposed method.

Future works include the followings: first, we are planning to examine whether the syntactic structures of the question sentence is useful for selecting appropriate keywords from the question sentence. Secondly, it is interesting to see whether the keyword selection method proposed in this paper is also effective for other applications such as answer candidate collection of the whole question answering process.

References

E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. 2002. Data-intensive question answering. In *Proc. TREC 2001*.

J. Fukumoto, T. Kato, and F. Masui. 2004. Question answering challenge for five ranked answers and list answers -overview of ntcir4 qac2 subtask 1 and 2-. In *Proc. 4th NTCIR Workshop Meeting*.

Sadao Kurohashi and Makoto Nagao, 1999. *Japanese Morphological Analysis System JUMAN version 3.62 Manual*.

Sadao Kurohashi, 1998. *Japanese Dependency/Case Structure Analyzer KNP version 2.0b6 Manual*.

C. C. T. Kwok, O. Etzioni, and D. S. Weld. 2001. Scaling question answering to the web. In *Proc. the 10th WWW Conf.*, pages 150–161.

B. Magnini, M. Negri, R. Prevete, and H. Tanev. 2002. Is it the right answer? exploiting web redundancy for answer validation. In *Proc. 40th ACL*, pages 425–432.

D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, and F. Lacatusu. 2003. Lcc tools for question answering. In *Proc. TREC 2002*.

S. Sato and Y. Sasaki. 2003. Automatic collection of related terms from the web. In *Proc. 41st ACL*, pages 121–124.

E. M. Voorhees. 2004. Overview of the trec 2003 question answering track. In *Proc. TREC 2003*.