Answering Yes-No Questions by Keyword Distribution: KJP System at NTCIR-11 RITEVal Task

Yoshinobu Kano
Shizuoka University
kano@inf.shizuoka.ac.jp

ABSTRACT
Textual entailment is normally regarded as a deeper analysis issue among other NLP techniques. Most textual entailment approaches employ deeper syntactic and semantic analyses. In contrast to such approaches, we used a simple, but fundamentally important, keyword based technique. Our system architecture was built on our observation that many of textual entailment issues are knowledge search issues, and extracted keyword distribution is the inevitable fundamental basis to solve the problem regardless of methods employed so far.

Keywords
RITE, textual entailment, keyword extraction, keyword distribution, knowledge search

Team Name
KJP

Subtasks
Japanese-Fact Validation (JA-FV).

1. INTRODUCTION
Textual entailment is normally regarded as a deeper analysis issue among other NLP techniques. Logic, reasoning, and deeper semantic analysis might be required if we aim to reach 100% performance. However, almost all of such techniques rely on keyword extraction. Given a proposition in the entailment task, we determine yes or no using knowledge sources. A first step of such a determination process is always keyword extraction, and then tries to perform deeper analysis based on the keyword extraction results. This hierarchical process implies that keyword distribution in knowledge sources would play a critical role.

In this paper, we assume that keyword distribution is sufficient to perform textual entailment tasks with a certain performance. Is this assumption sound too simple? There are two issues that are simple but not easy nor trivial. Firstly, as we described above, almost all of techniques rely on keyword extraction results due to the composite nature of modern NLP techniques. Therefore, poor keyword handling will spoil the entire system performance even if deeper techniques used. Secondly, researchers in the NLP community tend not to investigate such a fundamental level of processing. This does not mean the performance of keyword extraction is saturated, as we describe in this paper.

2. PREVIOUS WORKS
The Todai-Robot project [1] aims to solve university entrance examinations automatically as a challenging task of artificial intelligence. The target examinations include the Center Test, which is the very problems used in this RITEVal task (Matsuyoshi et al., 2014). We participated the Todai-Robot project and performed the best result for the History subjects in the Yozemi Mock Exam Challenge 2013 (Kano, 2014).

Our system is almost same as the one used in the Yozemi Mock Exam Challenge 2013. Thus, the system architecture and parameters were originally tuned for the real settings of the Center Test examination imitating like human applicants.

3. RITEVal
RITEVal's dataset was developed from the past Japanese National Center Test questions for University Admissions (Center Test). The Center Test is the common examination for Japanese students when applying to universities. The Center Test asks students multiple-choice style questions. The RITEVal dataset consists of three types of questions, "select the correct choice" type, "select the wrong choice" type, and "combination" type.

In the RITEVal task, the original multiple-choices were not given as a whole, but given one by one. In "select the correct choice" type questions, given a choice, RITEVal participant systems are asked to return a confidence value for that choice. Evaluation is performed by comparing confidence values for each original multiple-choices, regarding the largest value as the participant system's answer (smallest in case of "select wrong choice" type questions). In the "combination" type questions, the system is required to label Y or N for each choice and evaluated by a combination of these Y/N w.r.t the original multiple-choice question. In this paper, we focus on the "select correct/wrong choice" type questions.

Figure 1 illustrates an example set of choices in the RITEVal dataset. In this example, one of the four choices is the correct one. We added English translations to the original Japanese sentences.

As shown in the figure, domain, location and age could be different in the choices of the same question. Thus, it is not clear for which domain the system should search for the knowledge source. In addition, the expressions used in the questions and those used in the knowledge source are usually different and may be described over several sentences. Furthermore, in the case of wrong choices, there should be no corresponding part in knowledge source. These observations demonstrate that this task is difficult for machines to solve.

4. SYSTEM
4.1 Knowledge Source
In the RITEVal task, preprocessed data of the whole Wikipedia text and high school textbook texts were provided. We only participated to the Japanese subtask where all of propositions are in Japanese. Although we only used Japanese knowledge sources,

[1] http://21robot.org/

1. ポルトガルは 12 世紀，神聖ローマ帝国から独立した。
   Portugal attained independence from the Holy Roman Empire in the 12th century.
2. スペイン国王カルロス 1 世は，ポルトガル王を兼ねた。
   The King of Spain Carlos I also hold the King of Portugal.
3. スペインの作家セルバンテスが，『ドン＝キホーテ』を著した。
   A Spanish writer Cervantes wrote “Don Quijote”.
4. グラナダに，ロココ様式を代表するアルハンブラ宮殿が建設された。
   At Granada, the Palace of Alhambra was built, which is the hallmark of the Rococo style.

Figure 1. An example of RITEVal style problem. No.3 is the correct answer in this example.

our system architecture is language independent as described in later sections. We simply call the RITEVal Japanese Fact Validation subtask as RITEVal in this paper.

Wikipedia is a typical web sourced knowledge source. However, we decided to use only the textbook data in our system. The reasons are as follows.

First, the questions in the RITEVal task were taken from the Center Test questions. Since the Center Test tries to measure how much the students can solve the questions learned in their high school, the questions are composed of knowledge that is learnable from the high school textbooks.

Second, the structure of high school textbook is clean. That is, textbook tends to use a single place (snippet) for a single topic. For example, in case of history textbooks, a single historical event tends to be described in a single place.

Third, the sentences of high school textbooks are usually affirmative sentences. In other words, negative expressions (e.g., “an event did not occur”) do not usually appear.

These observations allow us to construct a high-precision system with simple techniques as described in the next section. These observations also answer the issue raised in the previous section, “where to search for knowledge source”. The answer is “searching the textbook data for the most relevant part using keyword distributions”.

4.2 Domain Independent Scoring

Our algorithm is based on the second and third observations. That is, we assume that the answer of a question is described in a single corresponding place (snippet) in textbooks with affirmative expressions.

We also design our system from the standpoint of domain independence. While the data we use in our system are small as described above, the high school textbooks have a lot of domains. In order to ensure that our system is applicable to different domains, we design our system to use an unsupervised method for QA.

Before describing the details of our system, we first recall the structure of the RITEVal task. As described in Section 2, an input to our QA system is a choice of a question extracted from the Center Test. Our QA system outputs a confidence score w.r.t. the input in the case of select correct/wrong type questions. In other words, let x be the given choice, our system output S(x) as the confidence of x.

Roughly speaking, our system performs (1) keyword extraction from the input, (2) keyword weighting of the input, (3) textbook search and scoring.

(1) Keyword extraction

Because Japanese texts are concatenation of characters not having spaces between words, we apply a morphological analyzer Kuromoji [2], which is based on Mecab [3], to the input. We augmented the dictionary of Kuromoji with the headings of the entries of Japanese Wikipedia. We extract all strings that match with the Wikipedia entries by longest match in the input as the keywords.

(2) Keyword weighting

Let c_i be the frequency of i-th distinct keyword in the input, then the weight of i-th keyword is

$$w_i = 1/(c_i z)$$

In this equation, $z = \sum_i 1/c_i$ is a normalizing constant, where i is defined over the distinct keywords in the input. The frequency c_i was counted over the textbook data.

(3) Textbook search and scoring

We divided the textbook data into snippets. We tried three types of snippets as described later. We search for the snippet that has the highest score w.r.t the input keyword set $\boldsymbol{K}$, which consists of the keywords in the input.

Let $\boldsymbol{R}$ be the word set extracted from a snippet, then the score of $\boldsymbol{R}$ is

$$s_R = \sum_{l \in R \cap K} w_l - \sum_{m \in K - R} w_m$$

This expression means that the score of the snippet is the sum of the weights of the input keywords included in the snippet minus that not included in the snippet. If a given choice is correct, keywords in the choice should be densely included in a specific snippet of the textbook; if a given choice is wrong, its keywords should be scattered across snippets. The above equation penalizes such a scattered keyword distribution.

[2] http://www.atilika.org/
[3] https://code.google.com/p/mecab/

Table 1. Evaluation results of our best submission. JA/JB, MS, PE, WA/WB stands for subjects of Japanese History A/B, Modern Society, Politics and Economics, World History A/B, respectively.

Macro-F1	Accuracy	JA	JB	MS	PE	WA	WB
57.00	57.59	0.579	0.375	0.250	0.280	0.286	0.174

Finally, we regard the maximum S_R among all of snippets as the confidence score of the corresponding input.

We do not consider negations because textbooks normally describe events in an affirmative way.

Our proposed method above does not depend on any domain specific information, even on any specific language.

5. EXPERIMENTS

Experiments were conducted on the RITEVal Japanese Fact Validation subtask dataset. All of our evaluation results are on the test data set using the RITEVal official evaluation tool. Since our system is unsupervised, we did not use the development set.

Table 1 shows results of our proposed method. As described in the previous section, we used three types of snippets, section, subsection and paragraph, larger to smaller in this order. These text sections were originally explicitly marked in the textbooks.

The RITEVal dataset was taken from various subjects including Japanese History A/B, Modern Society, Politics and Economics, and World History A/B. We have corresponding textbooks for each subject. However, in the RITEVal task, no information was provided about the original subject of a given proposition. Therefore, we used all the textbooks combined as a large single textbook when searching for the corresponding snippet. These textbooks include textbooks of Yamakawa Shuppansha and Tokyo Shoseki.

6. DISCUSSION

Our system is originally designed to solve the real problems in the Center Test. Although the RITEVal data set is created from the very Center Test, there are a couple of differences form the original problems.

Firstly, the original subject is not given for each proposition in RITEVal. This is a large disadvantage for our method because combining textbooks of different subjects will increase potential ambiguity due to overlapping keywords across subjects when finding corresponding snippets, decreasing the performance regardless of the subjects.

Secondly, there could be certain bias when selecting RITEVal propositions from the original Center Test problems. We observed that our system performance is better in the order of World History, Japanese History, Politics and Economics in the evaluation of all of available years of past Center Test problems without any modification nor selection for which problem to solve. A reason is that we tuned our system targeting at World History. Another reason may be the worse subjects require more "common sense" that are not explicitly described in textbooks. In contrast to these past observations, our RITEVal result shows that World History is the worst among the subjects. We obtained around 50% correct answer ratio in World History throughout past available Center Test problems while our score of World History B is lower than the chance level. Because the number of the past available problems we evaluated are larger than RITEVal's ones, there could be certain bias in problem selection.

Thirdly, the original Center Test problems are described in scattered paragraphs. These paragraphs were manually converted into a single sentence for each proposition of RITEVal dataset. In the original Center Test, finding relevant parts of paragraphs to extract keywords is one of the critical issues. This is because some parts of paragraphs or sentences are not directly related to solve the given problem. This difference may have led our relatively lower score compared to the scores for the original Center Test style problems.

7. CONCLUSION

Our keyword distribution based method achieved a state-of-art level score in RITEVal Japanese Fact Validation subtask. This result is, however, lower than the results we evaluated in the past Center Test and the Yozemi Mock Exam challenge. This difference would reflect differences between the original Center Test and the RITEVal dataset, such as problem selection bias, original subject specification, paragraph re-formation, etc.

Because our method is language-independent, domain-independent and unsupervised, our system could be used in other purposes by replacing the knowledge sources.

Using deeper analyses such as dependencies based on the current keyword based distribution would be a future work.

8. ACKNOWLEDGMENTS

This work was partially supported by JST PRESTO and the Todai Robot Project. We wish to thank Yamakawa Shuppansha and Tokyo Shoseki for providing there electronic textbook data.

9. REFERENCES

Kano, Y. (2014). Solving History Problems of the National Center Test for University Admissions (in Japanese). *Proceedings of the Annual Conference of JSAI*.

Matsuyoshi, S., Miyao, Y., Shibata, T., Lin, C.-J., Shih, C.-W., Watanabe, Y., & Mitamura, T. (2014). Overview of the NTCIR-11 Recognizing Inference in TExt and Validation (RITE-VAL) Task. *Proceedings of the 11th NTCIR Conference*.