A System to Solve Language Tests for Second Grade Students

Manami Saito
Nagaoka University of Technology
saito@nlp.nagaokaut.ac.jp

Kazuhide Yamamoto
Nagaoka University of Technology
yamamoto@fw.ipsj.or.jp

Satoshi Sekine
New York University
Language Craft
sekine@cs.nyu.edu

Hitoshi Isahara
National Institute of Information and Communications Technology
isahara@nict.go.jp

Abstract

This paper describes a system which solves language tests for second grade students (7 years old). In Japan, there are materials for students to measure understanding of what they studied, just like SAT for high school students in US. We use textbooks for the students as the target material of this study. Questions in the materials are classified into four types: questions about Chinese character (Kanji), about word knowledge, reading comprehension, and composition. This program doesn't resolve the composition and some other questions which are not easy to be implemented in text forms. We built a subsystem for each finer type of questions. As a result, we achieved 55% - 83% accuracy in answering questions in unseen materials.

1 Introduction

This paper describes a system which solves language tests for second grade students (7 years old). We have the following two objections.

First, we aim to realize the NLP technologies into the form which can be easily observed by ordinary people. It is difficult to evaluate NLP technology clearly by ordinary people. Thus, we set the target to answer second grade Japanese language test, as an example of intelligible application to ordinary people. The ability of this program will be shown by scores which are familiar to ordinary people.

Second aim is to observe the problems of the NLP technologies by degrading the level of target materials. Those of the current NLP research are usually difficult, such as newspapers or technological texts. They require high accuracy language processing, complex world knowledge or semantic processing. The NLP problems would become more apparent when we degrade the target materials. Although questions for second grade students also require world knowledge, it is expected that the questions become simpler and are resolved without tangled techniques.

2 Related Works

Hirschman et al. (1999) and Charniak et al. (2000) proposed systems to solve "Reading Comprehension." Hirschman et al. (1999) developed "Deep Read," which is a system to select sentences in the text which include answers to a question. In their experiments, the types of questions are limited to "When," "Who" and so on. The system is basically an information retrieval system which selects a sentence, instead of a document, based on the bag-of-words method. That system retrieves the sentence containing the answer at 30-40% of the time on the tests of third to sixth grade materials. In short, Deep Read is very restricted compared to our system. Charniak et al. (2000) built a system improved over the Deep Read by giving more weights for verb and subject, and introduced heuristic rules for "Why" question. Though, the essential target and method are the same as that of Deep Read.

3 Question Classification

First, we bought five language test books for second grade students and one of them, published by KUMON, was used as a training text to develop our system. The other four books are referred occasionally. Second, we classified the questions in the training text into four types: questions about Chinese character (Kanji), questions on word knowledge, reading comprehension, and composition. We will call these types as major types. Each of the "major types" is classified into several "minor types." Table 1 shows four major types and their minor types. In practice, each minor type farther has different style of questions; such as description question, choice question, and true-false question. The questions can be classified into approximately 100 categories. We observed that some questions in other books are mostly similar; however there are several questions which are not covered by the categories.

Major type	**Minor type**
Kanji	Reading, Writing, Radical, The order of writing, Classification
Word knowledge	Katakana, How to use Kana, Appropriate Noun, To fill blanks for Verb, Adjective, and Adjunct, Synonym, Antonym, Particle, Conjunction, Onomatopoeia, Polite Expression, Punctuation mark
Reading comprehension	Who, What, When, Where, How, Why question, Extract specific phrases, Progress order of a story, Prose and Verse
Composition	Constructing sentence, How to write composition

Table 1. Question types

4 Answering questions and evaluation result

In this section, we describe the programs to solve the questions for each of the 100 categories and these evaluation results. First we classified questions. Some questions are difficult to cover by the system such as the stroke order of writing Kanji. For about 90% of all the question types other than such questions, we created programs for basically one or two categories of questions. There are 47 programs for the categories found in the training data.

In each section of 4.1 to 4.3, we describe how to solve questions of typical types and the evaluation results. The evaluation results of the total system will be reported in the following section.

Table 2 to 4 show the distributions of minor types in each major type, "Kanji," "Word knowledge," and "Reading comprehension," in the training data and the evaluation results. Training and evaluation data have no overlap.

4.1 Kanji questions

Most of the Kanji questions are categorized into "reading" and "writing." Morphological analysis is used in the questions of both reading and writing Kanji; we found that large corpus is effective to choose the answer from Kanji candidates given from dictionary. Table 2 shows it in detail. This system cannot answer to the questions which are asking the order of writing Kanji, because it is difficult to put it into digital format.

The system made 7 errors out of 334 questions. The most of the questions are the errors in reading Kanji by morphological analysis.

In particular, morphological analysis is the only effective method to answer questions on this type. It would be very helpful, if we had a large corpus considering reading information, but there is no such corpus.

Question type	The rate of Q in training data[%]	The used knowledge and tools	Training data	Test data
			Correct ans. (total)	Correct ans. (known type Q., total)
Reading	27	Kanji dictionary, Morphological analysis	96(100)	6(8,8)
Writing	61	Word dictionary, Large corpus	220(222)	63(66,66)
Order of writing	6	-	0(20)	0(0,2)
Combination of Kanji parts	3	-	0(10)	0(0,0)
Classify Kanji	3	Word dictionary, Thesaurus	11(12)	0(0,0)
Total	100	-	327(364)	69(74,76)

Table 2. Question types for Kanji

4.2 Word knowledge questions

The word knowledge question dealt with vocabulary, different kinds of words and the structure of sentence. These don't include Kanji questions and reading comprehension. Table 3 shows different types of questions in this type.

For the questions on antonym, the system can answer correctly by choosing most relevant answer candidate using the large corpus out of multiple candidates found in the antonym dictionary.

The questions about synonyms ask relations of priority/inferiority between words and choosing the word in a different group. These questions can usually be answered using thesaurus.

Ex.1 shows a question about particle, Japanese postposition, which asks to select the most appropriate particle for the sentence.

The system produces all possible sentences with the particle choices, and finds most likely sentences in a corpus. In Ex.1, all combinations are not in a corpus, therefore shorter parts of the sentence are used to find in the corpus (e.g. "りんご (1) 買う。", "みかん (2) 買う。"). In this case, the most frequent particle for (1) is "を" in a corpus, so this system outputs incorrect answer.

Ex.1 [を/と/に]のうち、()に合うことばを書きなさい。
Select particle which fits the sentence from {wo,to,ni}
[1] りんご (1) みかん (2) 買う。
apple-(1) orange-(2) buy
(1) correct=と (to) system=を (wo)
(2) correct=を (wo) system=を (wo)

The questions of Katakana can be answered mostly by the dictionary. The accuracy of this type is not so high, found in Table 2, because there are questions asking the origin of the words, most Katakana words in Japanese has a few origins: borrowed words, onomatopoeia, and others. Because we don't have such knowledge, we could not answer those questions.

The questions of onomatopoeia include those shown in Ex.2. The system uses co-occurrence of words in the given sentence and each answer candidate to choose the correct answer in Ex.2, "ごろごろ." However, it was not chosen because the co-occurrence frequency of "ゆっくり," the word in the sentence, and "のろのろ," incorrect answer, is higher.

Ex.2 つぎの ようすを あらわす ことばを [ ] から えらんで、○で かこみなさい。
Choose the most appropriate onomatopoeia
(1) 大きな ものが ゆっくり ころがるようす。(A large object is rowing slowly)
[ころころ・ごろごろ・のろのろ]

The questions of word knowledge are classified into 29 types. We made a subsystem for each type. As there are possibly more types in other books, making a subsystem for each type is very costly. One of the future directions of this study is to solve this problem.

Question type	The rate of Q in training data[%]	The used knowledge and tools	Training data	Test data
			Correct ans. (total)	Correct ans. (known type Q., total)
Anonym	18	Antonym dictionary, Large corpus	26(27)	12(15,21)
Synonym	11	Thesaurus	14(17)	34(44,83)
Particle	19	Large corpus	25(28)	16(17,17)
Katakana	25	Word dictionary, Morphological analysis	18(37)	19(22,52)
Onomatopoeia	19	Large corpus, Morphological analysis	18(29)	16(20,31)
Structure of sentence	5	Morphological analysis	7(7)	20(22,22)
How to use kana	2	-	0(3)	0(0,19)
Dictation of verb	2	-	0(3)	0(0,0)
Total	100	-	108(151)	117(140,245)

Table 3. Question types for Word knowledge

4.3 Reading comprehension questions

The reading comprehension questions need to read a story and answer questions on the story. We will describe five typical techniques that are used at different types of questions, shown in Table 4.

Pattern matching (a) is used in questions to fill blanks in an expression which describes a part of the story. In general, the sequence of word used in the matching is the entire expression, but if no match was found, smaller portions are used in the matching.

Ex.3 Fill blanks in the expression
Story（partial）：「二、三日 たつと、その 花は しぼんで、だんだん 黒っぽい 色に かわって いきます。」
(In a few days, the flower withers and gradually changes its color to black.)
Expression：「花は (1)、(2) 色に かわる。」
(The flower (1) and change its color to (2).)
Answer：(1) しぼんで (withers)
(2) 黒っぽい (black)

The effectiveness of this technique is found in this example. The other methods will be needed when questions will be more difficult. At the time, this technique is very useful to solve many questions in reading comprehension.

For example, when questions start with "When (いつ)" and "Where (どこ)," we can restrict the type of answer word to time or location, respectively. If the question includes the expression of "どこに" ("に" is a particle to indicate direction/specification), the answer is also likely to be expressed with "ni" right after the location in the story. (**The kind of NE (Named Entity) and particle right after word (b)**)

For the questions asking the time or location about the entire story, this system outputs the appropriate type of the word which appeared first in the story. Although there are mistakes due to morphological analysis and NE extraction, this technique is also consistently very useful.

The technique which is **partial matching with keywords (c)** is used to seek an answer from story for "how," "why" or "of what" questions. Keywords from the question are used to locate the answer in the story.

Ex.4 かえったばかりの ひなは、どれくらいの 大きさですか。(How big are chicks when they hatched?)

Text（partial）：「かえったばかりの ひなは、おやゆびの先ぐらいの 大きさです。」(The size of chicks when they hatched is about the size of your thumb.)

Answer：おやゆびの 先ぐらい (size of thumb)

When the question is "Why" question, keywords such as "それで (thus)" and "だから (because)" are used.

Frequency in the large corpus is used to find the appropriate sentence conjunction. (d) Answer is chosen by comparing the mutual information of the candidate conjunctions and the last basic block of the previous sentence. However, this technique turns out to be not effective. Discourse analysis considering wider context is required to solve this question.

The technique which uses **distance between keywords in question and answers (e)** is supplementary to the abovementioned methods. If multiple answers are found, the answer candidate that is the closest in the story text to the keywords in questions is generated. These keywords are content words and unknown words in the text. This technique is found very effective.

In Table 4, the column "Used method" shows the techniques used to solve the type of questions, in the order of priority. "f" in the table denotes means that we use a method which was not mentioned above.

Question type		The rate of questions in training data [%]	Used methods	Training data corrent wns. (total)	Test data corrent ans. (known type Q, total)
Who	Who said	5	b,a,f	5(6)	1(4, 4)
	The others	0	b,e,f		
What	Like what	22	c,b,e	17(26)	1(1, 6)
	Of what		c,f,e		
	What doing		c,a,f		
	What is		a,e		
	What do A say		b,a,f,e		
When	Whole story	4	b	3(5)	0(0, 0)
	Part of story		-		
Where	Whole story	4	b	3(5)	0(1, 1)
	Part of story		b,f,c		
Why		16	c,f	11(18)	0(1, 1)
How		10	c	8(11)	0(0, 1)
How long, how often, how large		2	b,c,f	1(2)	0(0, 0)
To fill blanks		10	a	10(12)	4(9, 9)
Not have interrogative pronoun		4	-	0(5)	0(0, 0)
Conjunction		2	d	1(2)	1(3, 3)
Progress order of a story		10	f	8(11)	0(0, 0)
Paragraph		10	f	7(12)	3(3, 3)
The others		1	-	0(1)	0(0, 6)
Total		100	-	74(116)	10(22, 34)

Table 4. Question types for Reading comprehension

5 Evaluation

We collected questions for the test from different books of the training data. The proposition of the number of questions for different sections is not the same as that of the training data. Table 2 to 4 show the evaluation results in the test data for each type. Table 5 shows the summary of the evaluation result. In the test, we use only the questions of the type in training data. The tables also show the total number of questions, the number of questions which are solved correctly, and the number of questions which are not one of the types the system targeted (not a type in the training data).

The ratio of the questions covered by the system, questions in test data which have the same type in the training data are 97.4% in Kanji, 57.1% in word knowledge, and 64.7% in reading comprehension. It indicates that about a half of the questions in word knowledge isn't covered. As the result, accuracy on the questions of the covered types in word knowledge is 83.6%, but it drops to 47.8% for the entire questions. It is because our system classified the questions into many small types and builds a subsystem for each type.

The accuracy for the questions of covered type is 83.4%. In particular, for the questions of Kanji and word knowledge, the scores in the test data are nearly the same as those in the training data. It presents that the accuracy of the system is provided that the question is in the covered type. However, the score of reading comprehension is lower in the test data. We believe that this is mainly due to the small test data of reading comprehension (only 34) and that the accuracy for "Who" questions and the questions to fill blanks in the test data are quite difficult compared to the training data.

	Num. of all Q	Num. of known type Q	Num. of corrent ans.	RCA * (known type Q) [%]	RCA* (total) [%]	RCA* in total of known type Q [%]
Kanji	76	74	69	93.2	90.8	89.8
Word knowledge	245	140	117	83.6	47.8	71.5
Reading Comprehension	34	22	10	45.5	29.4	63.8
Total	355	235	196	83.4	55.2	80.3

Table 5. Evaluations at test data

* Rate of Correct Answer

6 Discussions

We will discuss the problems and future directions found by the experiments.

6.1 Recognition and Classification of Questions

In order to solve a question in language test, students have to recognize the type of the question. The current system skips this process. In this system, we set up about 100 types of questions referring the training data and a subprogram solves questions corresponding to each type. There are two problems to be solved. First, we have to design the appropriate classification and avoid unknown types in the test data. From the experiment, we found that the current types are not enough to solve this problem. Second, the program has to classify the questions automatically. We are building this system and are forecasting it quite optimistically once a good format is provided.

6.2 Effectiveness of Large Corpus

The large corpus of newspapers and the Web are used effectively in many different cases. We will describe several examples.

In Japanese, there are different Kanji for the same reading. For example, Kanji for "あう(au: to see, to solve correctly) are "会う(to see)" or "合う(to solve correctly)" for "人にあう(to see people)" and "答えがあう(to solve an answer correctly)," respectively. This type of questions can be solved by counting the expressions with Kanji in the corpus. It is similar to word sense disambiguation.

In the questions of particle complement, such as "かさ (umbrella) {に/と/を (locative-, conjunctive-, and objective particles)} 家 (home) {に/と/を} わすれる (to left) 。 (Intentional sentence is 'I left the umbrella at home')", it can be solved by counting the expressions with each particle in a corpus. This method is mentioned in Matsui（2004）but the evaluation result was not reported. When the answer is not found for the entire expression, the answer is searched by deleting some contexts. Most questions of filling blank types, similar strategy is helpful to find the correct answer.

In summary, the experiments showed that the large corpus is quite useful in several types of

questions. We believe it would be quite difficult to achieve the same accuracy by compiled knowledge, such as a dictionary of verbs, antonyms, synonyms, and relation words, and a thesaurus.

6.3 World Knowledge

The questions sometimes need various types of world knowledge. For example, "A student enters junior high school after graduated from elementary school." And "People become happy, if he receives something nice from someone." It is a difficult problem how to describe and how use that knowledge. Another type of world knowledge includes origin of words, such as foreign borrowed word or onomatopoeia. As far as we know, there is no comprehensive knowledge of such in electronic form. It is required to design attributes of world knowledge and to use them flexibly when applying then to solve the questions.

6.4 Difference between Reading Comprehension and Question Answering

The current QA systems identify the NE type of questions and seek the answer candidate of the type. However, the questions in the reading comprehension don't limit the answer types to person and organization, even if the question is "Who" type question. For example, "raccoon dog behind our house" or "the moon" can be the answer. Also, the answer is not always a noun phrase, but can be a clause, for example, "the time when new leaves growing on a branch" for questions asking time. There are different kinds of questions, which are asking not the time of specific event but the time or season of the entire story. For example "When is this story about?" In this case, the question can't be answered by just extracting a noun phrase.

However, at the moment, we can't conclude if the question can or cannot be answered without really understanding it. Sometime, we can find a correct answer without reading the story down the line or understanding the story perfectly. It is one of the future works.

6.5 Other techniques: discourse and anaphora

Some techniques other than morphological analysis, frequency of appearance in a corpus, and question answering methods are used in our system. We raise two issues. One of those is the discourse analysis. It is required in the questions to assign the order of paragraphs, and to select appropriate sentence conjunction. The other is anaphora analysis, which is very important, not only to indicate the antecedent, but also to find the link of mentions of entities.

7 Conclusion

We develop a system to solve questions of second grade language tests. Our objectives are to demonstrate the NLP technologies to ordinary people, and to observe the problems of NLP by degrading the level of target materials. We achieved 55% - 83% accuracy and several interesting NLP problems were found.

References

Hirschman, L., Light, M., Breck, E. and Burger, J. D. "Deep READ: a Reading Comprehension system". ACL, 1999, pp 325-332.

Charniak et al., "Reading Comprehension Programs in a Statistical-language-Processing Class". Workshop on Reading Comprehension Tests as Evaluation for Computer-based Language Understanding Systems. 2000.

K. Matsui: "Search Technologies on WWW which utilize search engines". (In Japanese) Journal of Japanese Language, February, 2004, pp 34-43.

K. Yoshihira, Y. Takeda, S. Sekine: "KWIC System on WEB documents", (In Japanese) 10th Annual Meeting of Natural Language Processing, 2004, pp 137-139.

Figure 1. A Snapshot of the system (http://languagecraft.jp/dennou/)