A Computational Method of Complexity of Questions on Contents of English Sentences and its Evaluation

Hidenobu KUNICHIKA, Minoru URUSHIMA, Tsukasa HIRASHIMA and Akira TAKEUCHI
Dept. of AI, Kyushu Institute of Technology, Japan

Abstract

This paper describes a computational method of complexity of a question for an adaptive question and answer function in an intelligent support system for English learning, and its evaluation. To realize adaptive question and answer, systems should generate questions depending on both educational intentions and learner's understanding state. For generating suitable questions for a learner automatically, systems must know the factors which influence difficulty of questions, and have ability to calculate difficulty. Difficulty is composed of a learner dependent part and an independent part. The former is evaluated by referring to a student model. The latter is defined by enumerating factors which influence complexity of questions. We present a definition of complexity of questions along learners' answering process. We also describe experimentation of comparing the complexity of questions calculated by computer with the complexity evaluated by human.

1. Introduction

QA (Question and Answer) about the contents of a story is widely used in language learning. QA in a target language is effective for acquiring practical skills because learners use plural language skills to answer questions, i.e. to grasp the contents of a story and question sentences and to compose answer sentences. When a teacher and a learner practice such QA, the teacher will give suitable questions for the learner. That is to say, the teacher will give an easier question to the learner who is doing badly and give a more difficult question to the learner who is doing well.

Some computer assisted language learning systems are equipped with test functions which ask about the contents of a story. Most of them, however, use questions prepared beforehand [4, 9], so teachers are burdened with preparing questions and answers for every learning material. And it is very difficult to prepare sufficient amount of questions and answers to correspond with various states of learners’ understanding.

The target of our study is to realize a QA function which provides suitable questions for each learner and tailored advice according to the learner’s answers. To achieve the target, we need the following functions: (1) to understand English sentences, (2) to generate various kinds of question sentences automatically, (3) to select suitable questions for each learner from a set of generated question sentences and (4) to analyze learners’ answer sentences and to diagnose errors. In the previous studies, we have proposed a method of analyzing stories and representing their meanings [7] for the function (1), a method of generating various kinds of questions from the syntactic structures and the meanings of stories [8] for the function (2), and a method of identifying syntactic errors [7] for the function (4).

In order to select suitable questions for each learner by the third function, we need a method to evaluate difficulty of questions. This paper proposes a computational method of the difficulty.

In the second chapter, we describe the outline of the QA function. Then in the third chapter, we propose a definition and a computational method of difficulty of questions along learners’ answering process. After that we present how our definition reflects humans’ judgments. And finally, our conclusions and discussions are presented.

2. The outline of the QA function

We assume that users of the QA function have studied the contents of a story and grammatical knowledge of English. Learning targets of QA are to master the use of grammatical knowledge and to train for conversation. In this chapter, we describe processing flow and information used for calculating difficulty of questions.

2.1 Processing flow

First, the QA function generates question sentences about the story as many as possible from syntactic and semantic information extracted by natural language processing. The query forms which the QA function treats are alternative questions and questions using interrogative pronouns. Next, the function calculates difficulty of each question for a learner, and then, selects a suitable and purposive question for the learner by referring to the difficulty of questions, educational intentions and information about a story.

Learner’s answers are analyzed by natural language processing, and syntactic and semantic information are extracted. The QA function also identifies syntactical errors. After that it judges the answers semantically by comparing semantic information of the answers with semantic information of the story. Results of the judgment are stored into a student model and then used for selecting the next question.

If the learner’s answer is incorrect, the QA function corrects the error intelligently by referring to both the student model and a teaching paradigm. For example, if the learner frequently uses a particular grammatical knowledge incorrectly, the function teaches the knowledge in detail.

2.2 Controlling difficulty of questions

In order to keep learners’ high motivation, it is important to give suitable questions for each learner, which are not too easy nor too difficult. The QA function selects suitable questions by not only considering the learning targets but also referring to difficulty of questions for a learner. For example, when a learning target is to master the use of grammatical knowledge and a learner has a poor record, the function selects questions which have low value of difficulty and use particular grammatical knowledge. And when a learning target is to train for conversation and a learner has a good
mainly focuses on the definition of complexity, because available complexity and the learner's understanding state, we state. We call the former "complexity" of questions. By understanding state. The other is the learner's understanding on only sentence structure and independent of a learner's difficulty are classified into two categories; one is dependent difficulty of answering questions depends on difficulty as arrows, so context such as referring relationship by pre-

Thus difficulty of answering questions depends on difficulty

The definition of difficulty of questions

1. Complexity of understanding the contents of sentences

A story consists of one or more paragraphs, and a paragraph consists of one or more sentences, so learners understand not only the content of each sentence separately but also context such as the outline of the story and relationship between events. Thus we define the complexity of understanding the meaning of each sentence in the story as the sum of "the complexity of a sentence" D1a and "the complexity of the paragraph which includes the sentence" D1b as follows.

\[ D1 = D1a + D1b \]

(a) Complexity of a sentence

The complexity of a sentence is defined as follows.

\[ D1a = w1a1 * f1a1 + w1a2 * f1a2 + w1a3 * f1a3 + w1a4 * f1a4 + w1a5 * f1a5 \]

"f"s are the factors of the complexity and "w"s are their weights.

Factors of complexity of a sentence are as follows.

fla1) The number of propositions: A proposition is the smallest unit of meaning and meaning of one sentence is represented by a set of propositions [5]. And the number of propositions in a sentence influences on complexity of understanding the contents of the sentence [6]. One proposition consists of one predicate such as a verb, adjective or adverb, and one or more variables such as nouns or noun phrases. The number of propositions is counted by referring to the syntactic tree.

fla2) Degree of conflict on identifying a referent: It is necessary to identify anaphoric relation in order to understand the contents of sentences. Thus the number of can-

3. The definition of difficulty of questions

To answer questions, learners execute some processes. Thus difficulty of answering questions depends on difficulty of tasks executing the processes. Factors that influence the difficulty are classified into two categories; one is dependent on only sentence structure and independent of a learner's understanding state. The other is the learner's understanding state. We call the former "complexity" of questions. By combining complexity and the learner's understanding state, we can evaluate difficulty of questions for each learner. This paper mainly focuses on the definition of complexity, because available information about an understanding state depends on implementation of student models and it may differ system by system, e.g. overlay models which express knowledge a learner has and buggy models which express incorrect knowledge in addition. So, we show an example of defining difficulty by using a student model of an intelligent English learning environment HELEN [7].

This chapter presents the definition of complexity of questions along answering process and the method to calculate complexity of questions, then explains customization of complexity to difficulty by using the student model. Finally, examples of the calculation are shown.

3.1 Complexity along the answering process

Process of answering questions consists of the following three subprocesses; understanding the contents of a story, understanding the content of a question sentence, and composing an answer sentence. The complexity of a question is represented by a triplet (D1, D2, D3). Each element corresponds to the complexity of each process. Factors of each complexity are enumerated by referring to results of some cognitive scientists' research work, but the influence of each factor on the complexity has not been cleared yet. So, we assume that each complexity is proportional to its factors, and each complexity is calculated by weighted sum of the factors. In this section, we enumerate factors in each answering process and describe a computational method for each factor.

2.3 Information to calculate difficulty

Our method of calculating difficulty of questions uses syntactic and semantic information generated by natural language processing. We have proposed a method of extracting syntactic and semantic information of stories based on DCG (Definite Clause Grammar). Semantic information is also important to judge semantic correctness of learners' answers. Here we briefly describe our representation method of syntactic and semantic information.

Syntactic information consists of a syntactic tree, which expresses parts of speech and modification relationships of words and phrases, and a feature structure, which expresses both grammatical functions of words and phrases and grammatical information such as sentence structure and idioms.

Semantic information consists of time and space information and information about verbs, nouns and modifiers. Fig. 1 shows the semantic information of the four sentences placed on the upper left-hand corner. One of features of the expression is that each piece of information is stored separately and relationships of correspondence are expressed by links shown as arrows, so context such as referring relationship by pronouns, or the time order of some events can be used easily.

Figure 1 An example of semantic information
didates for a referent influences on complexity of the process [1]. We treat pronouns and words which follow definite articles as anaphoric expression. Our method counts the number of candidates for a referent except for the real referent by referring to the semantic information.

f1a3) Distance between anaphoric expression and its referent: Complexity of identifying anaphoric relation is also influenced by distance between anaphoric expression and its referent [1]. Our method uses the number of preposition between anaphoric expression and its referent as the distance. However, if another anaphoric expression which refers to the same referent appears between the target anaphoric expression and its referent, the complexity is reduced because of the effect of recalling the referent. Thus we define the distance as follows.

\[ D_{1b} = \frac{w_{1b1} \cdot f_{1b1}}{(f_{1a3} of the nearest anaphoric expression which refers the same referent / 2)} + \text{the number of propositions between the target anaphoric expression and its referent or the nearest anaphoric expression if the expression exists} \]

Our method searches the semantic information for anaphoric relation and calculates the distance by referring to the syntactic trees.

f1a4) Complexity of modifying relation: Complexity of modifying relation of words influences complexity of understanding the content of a sentence [2]. We assume that it is complex to understand the content of a sentence in which modifying relation exists across some words. Our method seeks the syntactic tree of the sentence for such modifying relation and counts the number of the relations.

f1a5) The number of clauses: When plural clauses exist in one sentence, it is necessary for understanding the content of a sentence to identify relationship between the events represented by the clauses. Our method calculates the complexity by counting the number of clauses in a sentence by referring to its syntactic tree.

(b) Complexity of a paragraph

The complexity of a paragraph is defined as follows.

\[ D_{1b} = w_{1b1} \cdot f_{1b1} \]

f1b1) Scatter of objects: The complexity of understanding the contents of sentences depends on the degree of coherence of propositions [5]. Generally, people think that there are some relationships between sentences if the same object appears in all of the sentences. Therefore, we focus on objects in sentences and use scatter of objects defined as follows as a factor on the degree of coherence.

\[ \text{Number of appearance of objects / (the number of different objects) - 1} \]

The number of appearance of objects means the total number of objects considering consecutiveness in their appearance, i.e. if one object appears in consecutive sentences, the number is 1. For example, when an object X appears in the first, second, fifth and eighth sentences and an object Y appears in the seventh and eighth sentences, the number of appearance of the object is 4. And the number of different objects is 2. So, the value of scatter of objects is 1.

(2) Complexity of understanding the content of a question sentence

Complexity of understanding the content of a question sentence is the same as the complexity of understanding a sentence. The formula is as follows.

\[ D_{2} = w_{21} \cdot f_{21} + w_{22} \cdot f_{22} + w_{23} \cdot f_{23} + w_{24} \cdot f_{24} + w_{25} \cdot f_{25} \]

f21) The number of propositions: The same as f1a1.

f22) Degree of conflict on identifying a referent: The same as f1a2.

f23) Distance between anaphoric expression and its referent: The same as f1a3.

f24) Complexity of modifying relation: The same as f1a4.

f25) The number of clauses: The same as f1a5.

(3) Complexity of composing an answer sentence

Generally, a variety of correct answers can be considered for each question. To calculate the complexity, we have to know the learner’s answer sentence before asking the question, but it is impossible. However, it is impossible before the learner answers a question. So, our method uses “Yes, ...” or “No, ...”, e.g. “Yes, he is.” for alternative questions and a sentence in a story used for question generation for questions using interrogative pronouns as an assumed answer. Complexity of composing an answer sentence for a question using an interrogative pronoun is the same as the complexity of understanding a sentence. The complexity of composing an answer sentence is defined as follows.

\[ D_{3} = w_{31} \cdot f_{31} + w_{32} \cdot f_{32} + w_{33} \cdot f_{33} + w_{34} \cdot f_{34} + w_{35} \cdot f_{35} + w_{36} \cdot f_{36} \]

f31) The query form: The complexity of composing an answer sentence is influenced by the query form of a question. To answer questions using interrogative pronouns is higher level because a learner must have the necessary writing skill to compose an answer sentence. When an answer follows “No” to alternative questions, a learner must have writing skill, but the learner would be able to compose an answer sentence by referring to the question sentence. So this case is easier than the previous case. We define that 2 and 1 are the values of this factor for interrogative pronouns and alternative questions, respectively.

When learners answer questions using interrogative pronouns, our method calculates the following factors in addition.

f32) The number of propositions: The same as f1a1.

f33) Degree of conflict on identifying a referent: The same as f1a2.

f34) Distance between anaphoric expression and its referent: The same as f1a3.

f35) Complexity of modifying relation: The same as f1a4.

f36) The number of clauses: The same as f1a5.
3.2 Customizing complexity

Difficulty of understanding the contents of sentences depends on learners’ knowledge [1]. So, in order to calculate difficulty of questions, it is necessary to refer to information about learners’ understanding states. It is very difficult to consider variations of student models here, so we present an example of customizing complexity of questions to difficulty by using a student model of HELEN generated by buggy rule.

HELEN has ability to assess learners’ understanding states about vocabulary and grammar. For example, HELEN judges that a learner has master use of a particular word. And HELEN also judges that a learner knows a particular grammatical structure but the learner can not use the structure well, that is to say, HELEN knows whether or not the structure is unfamiliar for the learner. HELEN represents influence of a learner’s understanding states as the following triplet.

\[ \text{D1} = w_1a_1 + w_1b_1 + w_1c_1 \]

\[ w_1a_1 = w_1a_1 + w_1a_2 + w_1a_3 + w_1a_4 \]

\[ w_1b_1 = w_1b_1 + w_1b_2 + w_1b_3 + w_1b_4 \]

\[ w_1c_1 = w_1c_1 + w_1c_2 + w_1c_3 + w_1c_4 \]

We present examples of complexity of questions about the junior high school students.

3.3 Examples of calculating complexity of questions

We present examples of complexity of questions about the contents of the following sentences which correspond to one paragraph in a textbook "New Horizon" [12] for Japanese junior high school students.

- He likes rock climbing.
- He goes to college in San Francisco.
- He is studying space science.
- This is my brother Fred.

The following question sentences are generated automatically from the above sentences by the automated question generation function [8]. The complexity of each question is represented by a triplet and each element of a triplet corresponds to the complexity of each answering process. In these examples, the weight of each factor is 1 and the influence of learners’ understanding states is not considered.

\[ (4, 2, 5) \text{ Where does Fred go?} \]

4. Evaluation

To evaluate the appropriateness of our definition of complexity, we compared the difficulty with humans’ judgment. We gave subjects two questions, and asked them to judge which is difficult. Then, we evaluated both the recall rate which expresses how our method covers humans’ judgment and the precision rate which expresses how our method corresponds to humans’ judgment.

In the previous chapter, we have enumerated factors which influence complexity of answering questions, and represented complexity of a question as a triplet which consists of complexity of understanding the contents of sentences, complexity of understanding the content of a question, and complexity of composing an answer sentence. A variety of ways to use these values can be considered, e.g. to adjust weights of factors depending on the educational intentions. In this evaluation, we define complexity of a question as the sum of three elements of the triplet.

4.1 The method of experimentation

First, we selected three stories in New Horizon and seven paragraphs in the stories at random. Then we generated question sentences from each paragraph by the automated question generation function. After that, we randomly selected four question sentences for each paragraph from the generated questions. We call the four question sentences "question set".

We gave paragraphs and question sets to 16 subjects who are undergraduate students, and asked the subjects to sort the question sentences of each question set in order of difficulty. We also asked them to write the reason of the decision. Because all sentences are gotten from the textbook for junior high school students and we have confirmed that there is no unfamiliar words and grammars for the subjects, the influence of subjects’ understanding states is excluded.

In this evaluation, we assumed that all factors affect equally to the complexity. That is to say, all weights of these factors are 1, and we calculated complexity of each question by adding up D1, D2 and D3. Because it is difficult to evaluate ordering of four questions by humans and our method at a time, we selected all combinations of two questions from a question set and compared judgments by humans and our method.

4.2 The results and consideration

(1) The recall rate

As judgment of all humans, in general, are not always the same, we compared judgment of our method with humans' judgment that most of the subjects; i.e. more than 13 subjects out of 16, got the same result, and we calculated the recall rate. Each question set consists of four questions, so

\[ \text{Recall rate} = \frac{\text{Number of correct judgments}}{\text{Total number of judgments}} \]

\[ \text{Precision rate} = \frac{\text{Number of correct judgments}}{\text{Total number of judgments}} \]
there are 6 pairs of questions in each question set. And there are 7 paragraphs. Thus there are 42 pairs of questions in total.

As the result, we have found that there are 23 pairs in which almost subjects have made the same judgment, and there are 19 pairs out of the 23 pairs in which our method have made the same judgment. Therefore, the recall rate is about 83%, and we assess that the complexity calculated by our method reflects the judgment by humans well. There are two reasons of disagreement of the judgments: one is that questions have no sufficient information to search answer and the other is that questions include relative pronouns.

(2) The precision rate

In order to select difficult or easy questions, the QA function compares values of difficulty of questions. Thus, it is necessary to define a threshold value used for judgment of significant difference. Moreover, in order to confirm validity of the judgment by using our method, it is necessary to evaluate the precision rate. We considered that the judgment that more than 13 subjects made the same is correct and calculated the precision rates by changing the threshold value from 1 to 10 by 1. Figure 2 shows the results. The bar graph in the figure shows each precision rate and the line graph shows each covering rate. A covering rate is a rate of number of the pairs in which difference of values of difficulty is bigger than the threshold value. We can recognize from the figure that when the threshold value is small, the validity of judgment is low but there are many available questions, and when the threshold value is big, there are a few available questions but the validity is high. Thus, we see that our method is useful if the threshold value is decided by considering the balance of the precision rate and the covering rate.

5. Conclusions and Discussions

This paper has described the definition of difficulty of questions for selecting adaptive questions for learners' understanding states and the mechanism of its calculation. We have compared judgments by humans and our method. As the result, we have seen that our method is useful if the threshold value is decided by considering the balance of the precision rate and the covering rate.

Difficulty of understanding sentences, in other words, objective measures of readability such as FOG [3], SMOG formula [10] and FORCAST formula [13] are widely used, but they mainly focus on the difficulty of words. Because users of the QA function are studying not only words but also grammatical knowledge, complexity of execute process of understanding the contents of sentences also strongly depends on grammatical factors. So it is not adequate to use only information of words for the QA function.

Some cognitive scientists try to clear the mechanism of natural language understanding. The mechanism has not been cleared yet, nevertheless some factors which influence complexity of the understanding are enumerated as their results. For example, Abe et al. [1] and Omura [11] present major factors. The factors used for calculating complexity of questions for our QA are chosen from the major factors by considering the object of our QA, that is to say, the QA function gives questions about the surface meaning of a story for novice learners. In order to realize the adaptive QA, it is necessary to calculate the complexity considering the factors automatically by computers. However, no mechanism of the calculation has been proposed, so we have defined the complexity along the answering process and proposed the mechanism of its calculation in this paper.

Next target of our research is to realize a function to select suitable questions by referring to difficulty of questions.

Acknowledgments

This work is supported in part by Grant-in-Aid for Scientific Research on Priority Areas (A) (1) No.12040104 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References