Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

# Interactive Environment for Learning by Problem-Posing of Arithmetic Word Problems Solved by One-step Multiplication and Division

**Sho YAMAMOTO[a*], Yuki AKAO[a], Mitsutaka MUROTSU[a], Takehiro KANBE[a], Yuta YOSHIDA[a], Kazushige MAEDA[b], Yusuke HAYASHI[a] & Tsukasa HIRASHIMA[a]**
[a]*Graduate School of Engineering, Hiroshima University, Japan*
[b]*Elementary School Attached to Hiroshima University, Japan*
*sho@lel.hiroshima-u.ac.jp

**Abstract:** Problem-posing is effective learning for comprehending problem structure. We have designed and developed the learning environment for problem-posing and performed its practical use for first and second grade students on elementary school continually. The scopes of these systems are one-step addition, subtraction or multiplication arithmetic word problem. The results of these practical uses suggested that the learning environment was effective to comprehend these problem structures. Therefore, as the next step, we have designed and developed a learning environment for posing one-step multiplication or division word problem in order to let learners acquire a difference between multiplication and division. Developed learning environment and its practical use in an elementary school are reported.

**Keywords:** Problem-posing, sentence-integration, interactive environment, multiplication, division, arithmetic word problem, problem structure

## 1. Introduction

Several researchers postulate problem-posing is effective exercise for promoting learners to master the use of solution methods (Polya, 1945; Silver, CAI, 1996). Moreover, it has been proposed that poor problem solvers often fail to elicit problem structures from problem (Mayer, 1982; Kintsch, Greeno, 1985). We design and develop the learning environment which learners acquiring the structure of arithmetic word problem by exercising the problem-posing. Now, our research domain is an arithmetic word problem can be solved by one-step calculation operation. We analyze a structure of arithmetic word problem and develop the learning environment based on its structure (Nakano, et al, 1999; Hirashima, et al, 2007; Hirashima, et al, 2011). Until now, one-step addition, subtraction or multiplication word problem are analyzed and the structure of these problems are implemented on tablet PC (Yamamoto, et al, 2012; Yamamoto, et al, 2013). In addition to the development, we have performed two experimental uses with elementary school teacher, which are first grade students by learning the one-step addition or subtraction and second grade students by learning the one-step addition or subtraction word problem. As this reason, the first grade students have learned the problem-posing by one-step addition or subtraction word problem because they have already known the concept of addition and subtraction in their life. The second grade students only learned to pose one-step multiplication problem because the concept of multiplication is a difficult concept for learners. The results of these experimental uses have proposed that not only the learner improve the problem solving performance, but also this learning environment was effective for the learner who can't judge the problem structure to acquire the problem structure.

The students are only required to consider the multiplication problem structure by learning the assignment of learning environment for second grade student, As the next step, the learner learn not only one-step multiplication but also one-step division word problem. Therefore, they are required to judge whether the story means one-step multiplication or division. For this learning, we have designed a assignment and developed the learning environment by problem-posing. In this paper, a problem structure is explained in the following chapter. A design of developed learning environment based on

this structure is described in section 3. A sequence of assignment is also explained. Subsequently, a procedure of its practical use and an analysis of the results are reported.

## 2. Problem Structure of One-step Multiplication or Division Word Problem

In this section, the model of one-step multiplication or division arithmetic word problem is explained. One-step arithmetic word problems can be expressed by three sentences in our research. Example is shown in Figure 1. Because there are three values in one-step arithmetic word problem, this problem can be expressed by three sentences. Each sentence consists of value, object and predicate. These sentences consist of two sentences mean existence and one sentence means relation between other two values. We call each sentence as existence sentence and relation sentence. In this example, "There are three boxes" and "There are several apples" are existence sentence. "There are five apples in each box" is relation sentence because this sentence shows the relation between the apple and box.

In addition to the kind of sentence, each sentence has a property of quantity in multiplication and division word problem (Yamamoto, et al, 2013). Generally, multiplication is expressed by "multiplicand multiplied by multiplier is product" (Greer, 1992; Vergnaud, 1983). So, it is said that each quantity has different property. This word problem contains the story that the value of apples is expressed as the amount of apple when there are three boxes and the value of apples in each box is basis. Since, in Japanese Education, multiplicand is also called "base quantity", multiplier is "proportion" and product is "compared quantity". Then, the arithmetic word problem that can be solved by one-step multiplication or division has three types of story. (1) Compared quantity divided by base quantity is proportion, (2) Base quantity multiplied by proportion is compared quantity, (3) Compered quantity divided by proportion is base quantity. The story of the problem in Figure 1 is (2).

All of these stories contain the relation that is “Base quantity multiplied by proportion is compared quantity”. One-step multiplication or division word problems are expressed by changing the one quantity to required value in each story. Therefore, it is important to extract the base quantity, proportion and compared quantity from problem and to make the relation between these quantities.

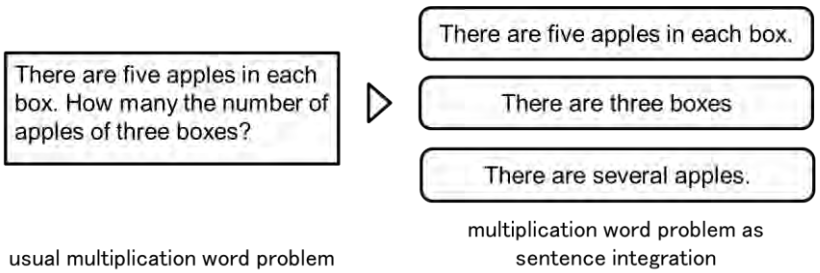


Figure 1. Example of Problem Expression as Sentence Integration.

## 3. Outline of Learning Environment for Problem-posing "MONSAKUN Touch3"

### *3.1 Framework*

This learning environment consists of MONSAKUN Touch 3 for learners and MONSAKUN Analyzer 3 for teachers. A result of the learner's learning by problem-posing on MONSAKUN Touch 3 is sent to database server via network. MONSAKUN Touch 3 developed by using Android, MONSAKUN Analyzer 3 by using PHP and JavaScript. Of course, the each software can be run on Android Tablet. RDBMS is used MySQL. The teacher can confirm the graph of learner’s learning by using MONSAKUN Analyzer 3 that receives a learning data from database server. The learning data are saved as three data: the number of correct problem, the number of incorrect problem, the number of the each incorrectness and the learner’s log. Category of incorrectness is based on a diagnosis of MONSAKUN Touch 3. MONSAKUN Analyzer 3 generates some graph by using these data and displays teacher it. Teacher can limit to an assignment that learner can exercise on MONSAKUN Touch 3 by using MONSAKUN Analyzer 3.

### *3.2 MONSAKUN Touch 3*

#### *3.2.1 Outline of MONSAKUN Touch 3*

In MONSAKUN Touch, after the learner logged in the environment and selected level, he/she sees an interface for problem-posing. This interface presents the assignment for posing problem, the set of given sentence card and three blank for arranging given sentence cards. The learner can pose the problem by selecting three sentence cards from given cards and arranging them in proper order. Given sentence cards are consists of correct card set and dummy card set for leading to errors. If three blank is filled with three sentence cards, diagnosis button will be active. Then, the learner can tap this button and the system diagnoses and generate a feedback his/her posed problem. When the learner finishes answering all assignment in selected level correctly, the interface for posing problem backs to the interface for selecting level. These flow and method of exercise are same as previous MONSAKUN. However, in MONSAKUN Touch 3, the text means the property of quantity is shown in the left side of each blank because let the learner consider the property of each sentence. The assignment that is described next section is renewed.

#### *3.2.2 Designing the Level of Assignment*

Table 1 shows the all level of assignment by dividing into the number of level, assignment, required activity, contents of assignment and number of assignment. Each level is designed so that the learner acquires to judge the structure of one-step multiplication or division word problem. The learner is required to pose the story from level 2 to 7, to pose the problem from level 8 to 9.

In level 1, the learner is given the story of one-step multiplication and four calculation expression which are expressed by "Base quantity multiplied by proportion is compared quantity", "Proportion multiplied by base quantity is compared quantity" and the cumulation of same number like “4+4+4=12” and “3+3+3+3=12”. This assignment is the confirmation of the relation between multiplication and addition. Then, the learner is required to select the correct calculation expression. The purpose of this level is which let the learner comprehend the relation of multiplication story and addition calculation. The learning environment gives the story and several sentence cards to the learner in level 2. The given story as sentence integration consists of two fixed sentence cards and one blank. The learner is required to fill this blank by considering the property of quantity. In this assignment, they learn the property of quantity that is contained each given sentence card. Given sentence cards in level 3 are included two sentence cards that have different text representation and same property. For example, "There are two boxes." and "The number of box is two.". In this level, let the learner learn that the sentence cards include the same property of quantity have various text representation. MONSAKUN Touch 3 present the three blank for putting the sentence cards and several sentence cards in level 4. Then, the learner is required to pose the story by selecting three sentence cards and by arranging them in proper order based on the relation of "Base quantity multiplied by proportion is compared quantity". The assignment of level 5 requires the learner to pose the two stories by using one common sentence card. Through this exercise, the learner comprehend that existence sentence card is able to have two property of quantity. In other words, both proportion and compared quantity are expressed by existence sentence. After that, in level 6, the learner learns that the story has three kinds of calculations expression that are mentioned in section 2. This purpose is that the learner notices the multiplication story contain the calculation (a) and (c). Thus, the learner is given the multiplication and division calculation expression as assignment for posing story. In order to let the learner confirm three properties of quantity and its relation again, assignments of level 7 includes improper assignment which cannot solve because of lack of one proper sentence card. Then, the learner is given a specific sentence card for posing the story in this level, which is labeled "proper sentence card is not given" instead of lacking sentence card. Because the assignments in level 7 are composed of usual assignment and assignment which mentioned above, the learner is required to consider each property of quantity and its relation again. As the next step, the learner is required to pose problem in level 8 because the learner learn to pose the story through level 2 to 7. Finally, in level 9, the learner is required to pose the two problems by using one common sentence card. This assignment is same as assignment of level 5. Through the exercise from level 1 to 9, the learner can acquire the problem structure gradually.

Table 1: The Assignment Level on MONSAKUN Touch 3.

| Level | Required activity | Contents of assignment | Number |
|---|---|---|---|
| 1 | Select calculation expression | Select calculation express given story | 12 |
| 2 | Pose story | Pose story that is expressed by given calculation (one-step multiplication) Required story has already given two sentence cards | 12 |
| 3 | Pose story | Same as assignment of level 2 Include same property and different text representation | 12 |
| 4 | Pose story | Pose story that is expressed by given calculation (one-step multiplication) Select three sentence cards and arrange them | 10 |
| 5 | Pose story | Pose two stories by using same sentence card | 10 |
| 6 | Pose story | Pose story that is expressed by given calculation But given calculation expression is one-step multiplication or division | 12 |
| 7 | Pose story | Same as assignment of level 6 But one proper sentence card is not given | 12 |
| 8 | Pose problem | Pose problem that is expressed by given calculation Select three sentence cards and arrange them | 12 |
| 9 | Pose problem | Pose two problems by using same sentence card | 12 |

### *3.3 Outline of MONSAKUN Analyzer 3*

After the teacher log in the learning environment by inputting id and password, MONSAKUN Analyzer displays the visualized learning data like Figure 2. These data are consists of the average score of the problem-posing and rate of each incorrectness in his/her each lesson. These data generates and shows as a three bar charts and a doughnut chart. Moreover, this environment indicates the average achievement of the level and assignment number. The count of posed problem and the rate of incorrectness are showed by not only each lesson but also each student. This function is same as previous system. In addition to this function, MONSAKUN Analyzer 3 can extract the data based on the each level and assignment from these data. For example, teacher can see the learning data of level 4. The progress of number of correct and incorrect posed problem is visualized by line chart.

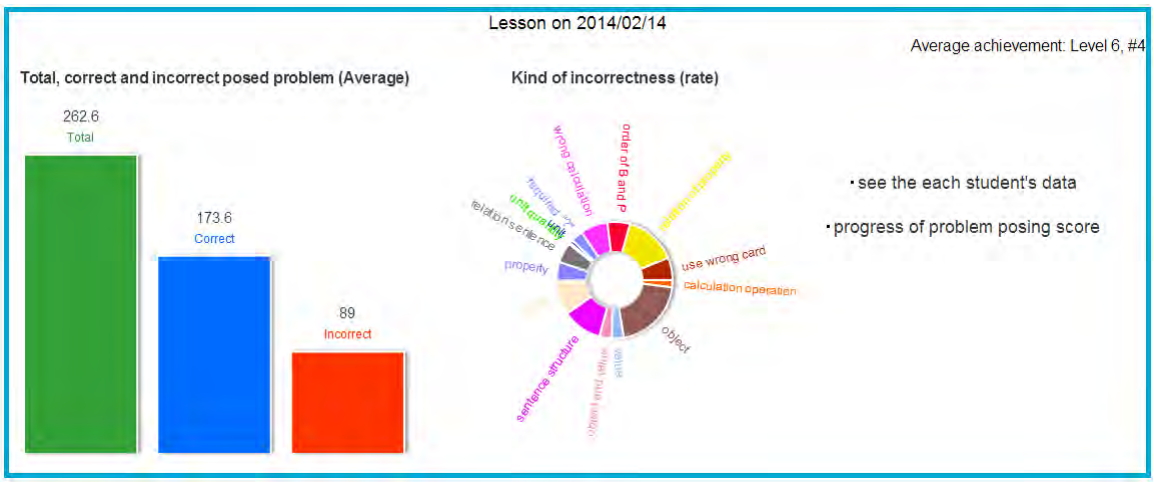


Figure 2. The Part of Main Interface of MONSAKUN Analyzer.

## 4. Experimental Use of MONSAKUN Touch 3

### *4.1 Procedure of Experimental Use*

The subjects were thirty-nine students in the third grade of an elementary school. They were divided into the subjects who experienced MONSAKUN Touch 1 and 2 in our previous research (Yamamoto, et al, 2012; Yamamoto, et al, 2013) and who did not experience it. Also, they had just learned to solve arithmetic word problems that can be solved by one-step multiplication or division. This experimental use has been performed during thirteen lessons that consist of pretest in one lesson, learning by MONSAKUN Touch 3 in eleven lessons and posttest in one lesson (45 minutes per lesson, in 5 weeks). A lesson by using MONSAKUN composes of teaching about problem-posing by a teacher and problem-posing exercise by using MONSAKUN Touch 3. The time of using MONSAKUN Touch 3 is decided by the teacher based on the progress of each lesson. If the subjects have finished twice the current level when they exercise the problem-posing after teaching, they were allowed to work on the previous level. The purpose of this experimental use is to examine the effects of the learning by MONSAKUN Touch 3 and the effects of experience MONSAKUN continually.

We used these three tests: usual problem solving test, extraneous problem solving test and problem-posing test. Usual problem solving test can be solved by one-step multiplication or division that is expressed by three sentences. Usual problem solving test has sixteen questions because each quantity can be the required value in these five stories. Extraneous problem solving test includes extraneous information that is not necessary to solve the problem (Muth, 1992). The extraneous problem solving test is useful to assess leaner's comprehension of the problem structure. These problems consists of twelve problems that including the two kinds of extraneous information that change sentence cards except sentence contains required value in each six stories. Problem-posing test examine the problem-posing performance to let the subject pose the problem as he/she can within the time limit. The subject pose problem from scratch. The time limit is ten minutes in each test. The difference between pretest and posttest is order of each problem.

### *4.2 Analysis of Pretest and Posttest*

Analysis of pretest and posttest are reported in this section. And the level by using lecture is described. The teacher performed the lecture based on the level on MONSAKUN Touch 3 and treated one level in one lecture. However, in level 3, it is difficult for the subjects to relate between multiplication calculation expression and text representation contain "cut" because "cut" is associated with division calculation expression. Thus, the teacher has to spend three lessons for resolving this difficulty.

The results of average score and SD in three tests are shown in Table 2. These scores are divided into experienced and inexperienced group of MONSAKUN in our previous research. In addition to this result, we analyze the result in each test by ANOVA. There was an interaction in the score of usual problem-posing test between experience of MONSAKUN and pre-posttest ($p$=.03). So, we analyzed simple effect. There was a significant difference in the score of posttest between experienced group and inexperienced group ($F$(1, 36)=3.193, $p$=.008). This result suggested that it is effective for the subjects to experience the learning by using MONSAKUN for improving their usual problem solving performance. Next, there was a significant difference in the score of extraneous problem solving test between experienced group and inexperienced group ($p$=.04), and effect size is medium ($|\eta^2|$=.10). Also, there was a significant difference in the score between pretest and posttest ($p$=.02), and effect size is small ($|\eta^2|$=.02). In addition to this analysis, we analyzed the correlation between the pretest score and the difference between posttest and posttest score. In this result, there are a negative correlation between them (Spearman's rank-correlation coefficient, $|rs|$=.59, $p$=2.5E-06). These results suggested that the lesson by using our learning environment promote the subjects to improve their problem structure, in particular, more effective to the subjects who have experienced MONSAKUN to comprehend the problem structure. MONSAKUN Touch 3 is more effective for the subjects who the score of extraneous problem solving test is lower particularly. Last, there was no significant difference in the number of posed problem between experienced and inexperienced group. But, there was a significant difference between pretest and posttest ($p$=.005), and effect size is medium ($|\eta^2|$=.07). These results suggested that MONSAKUN is useful for the subjects to improve their problem-posing performance regardless of whether the subjects have experienced MONSAKUN.

Table 2: Result of Each Pretest and Posttest in experienced group (N=18) and inexperienced (N=20).

| Test | Experience of MONSAKUN | Pretest | | Posttest | |
|---|---|---|---|---|---|
| | | M | SD | M | SD |
| Problem-posing | experienced | 2.50 | 1.57 | 3.56 | 1.42 |
| | inexperienced | 2.10 | 1.45 | 2.7 | 1.52 |
| Usual Problem solving | experienced | 13.61 | 1.34 | 14.28 | 0.80 |
| | inexperienced | 13.30 | 1.52 | 12.75 | 2.05 |
| Extraneous Problem Solving | experienced | 10.83 | 2.14 | 11.38 | 0.76 |
| | inexperienced | 8.80 | 3.54 | 9.75 | 3.40 |

## 5. Conclusion

In this paper, we have described the learning environment for problem-posing in one-step multiplication or division arithmetic word problem and its practical use. In order to realize the learning environment, firstly, we have mentioned that three quantities and its relation define one-step multiplication or division word problem. These three quantities are called base quantity, proportion and compared quantity. Its relation expresses the story that is "Base quantity multiplied by proportion is compared quantity" and so on. At the second step, the problem-posing based on this problem structure and the diagnosis and feedback of posed problem are defined. The levels of assignment were designed so that the learner can judge the story means whether multiplication or division. After that, we have developed learning environment called MONSAKUN Touch 3 and MONSAKUN Analyzer. Lastly, an eleven lesson experimental use is reported. The results of brief analysis suggested that the third grade students who have learned by using MONSAKUN in the past time are improved their problem solving performance and sophisticated their acquired problem structure.

As our future works, we need to verify the quantity of the effect to high group by using MONSAKUN Touch 3. Furthermore, we should perform the practical use for problem-posing in one-step addition, subtraction, multiplication or division to fourth grade students of an elementary school continuously.

## References


E.A. Silver, J. CAI. (1996). An analysis of arithmetic problem posing by middle school students. Journal for Research in Mathematics Education, 27(5), 521-539.

Greer, B. (1992). Multiplication and Division as Models of Situations. Handbook of Research on Mathematics Teaching and Learning, 276-295.

Hirashima, T., Kurayama, M. (2011). Learning by Problem-Posing for Reverse-Thinking Problems, Proc. of AIED 2011, 123-130.

Hirashima, T., Yokoyama, T. Okamoto, M. ,Takeuchi, A. (2007). Learning by Problem-Posing as Sentence-Integration and Experimental Use. Proc. of AIED2007, 254-261.

Kintsch, W., Greeno, J.G. (1985). Understanding and Solving Word Arithmetic Problem. Psychological Review, 92(1), 109-129.

Mayer, R. E. (1982). Memory for algebra story problem. Journal of Educational Psychology, 74, 199-216.

Muth K.D. (1992). Extraneous information and extra steps in arithmetic word problems, Contemporary educational psychology. Vol. 17, pp.278-285.

Nakano, A., Hirashima, T., Takeuchi, A. (1999). Problem-Making Practice to Master Solution-Methods in Intelligent Learning Environment. Proc. of ICCE'99, 891-898.

Polya, G. (1945). How to Solve It. Princeton University Press.

Vergnaud, G. (1983). Multiplicative Structures. Acquisition of mathematics concepts and processes, 127-174.

Yamamoto, S., Kanbe, T., Yoshida, Y., Maeda, K., Hirashima, T. (2012). A Case Study of Learning by Problem-Posing in Introductory Phase of Arithmetic Word Problems. Proc. of ICCE2012, Main Conference E-Book, 25-32.

Yamamoto, S., Takuya H., Kanbe, T., Yoshida, Y., Maeda, K., Hirashima, T. (2013). Interactive Environment for Learning by Problem-Posing of Arithmetic Word Problems Solved by One-step Multiplication, Proc. of ICCE2013, Main Conference E-Book, 51-60.

# Question Generation Using WordNet

**Nguyen-Thinh LE & Niels PINKWART**
*Humboldt-Universität zu Berlin, Germany*
{nguyen-thinh.le, niels.pinkwart}@hu-berlin.de

**Abstract:** Discourse and argumentation are effective techniques for education not only in social domains but also in science domains. However, it is difficult for some teachers to stimulate an active discussion between students because several students might not be able to develop their arguments. In this paper, we propose to use WordNet as a semantic source in order to generate questions that are intended to stimulate students brainstorming and to help them develop arguments in an argumentation process. In a study, we demonstrate that the system-generated questions sound naturally as human-generated questions as measured by computer scientists.

**Keywords:** Question generation, WordNet, argumentation

## 1. Introduction

Studies have reported that deploying questions are effective for learning. Asking targeted, specific questions is useful for revealing knowledge gaps with novices, who are often unable to articulate their questions (Tenenberg & Murphy, 2005). Other researchers used prompts as a kind of questions in order to encourage students to self-explain and demonstrated that prompts are a promising instructional feature to foster conceptual understanding (Berthold et al., 2011).

Argumentation is an important skill that is required in any situation, either in research or in daily life, and thus needs to be trained. In order to train students, usually, they are asked to discuss together about a given topic. That is, they need to develop arguments during the argumentation process. However, students may sometimes not proceed with their argumentation. In this paper, we propose to use questions in order to stimulate their brainstorming and the goal is that they use the posed questions to develop new arguments for a given discussion topic. How can questions that are semantically related to a given discussion topic be generated in order to help students develop further arguments?

In this paper, we introduce an approach to exploiting WordNet to generate questions which are related to a discussion topic and investigate the research question: *Do automatic system-generated questions appear as natural as human-generated questions?* This paper reports on results of an evaluation study that is intended to test the specified research question.

## 2. State of the Art of Question Generation for Educational Purposes

Traditionally, questions are generated from a text or from structured data and natural processing techniques are used to analyze a text and to construct a question. In the state of the art, Le and colleagues (Le et al., 2014) classified educational applications of automatic question generation into three classes according to their educational purposes: 1) knowledge/skills acquisition, 2) knowledge assessment, and 3) educational systems that use questions to provide tutorial dialogues.

Examples of the first class of educational applications of automatic question generation include the work of Kunichika et al. (2001) who extracted syntactic and semantic information from an original text and generated questions based on extracted information, the reading tutor of Mostow and Chen (2009), and the system G-Asks (Liu et al., 2012) for improving students' writing skills (e.g., citing sources to support arguments, presenting evidence in a persuasive manner). The second class of educational applications of question generation aims at assessing knowledge of students and includes the approach of Heilman and Smith (2010) for generating questions for assessing students' acquisition of factual knowledge from reading materials, the computer-aided environment for generating multiple-choice test items of Mitkov et al. (2006), and the REAP system of Brown et al (2005), intended to assess the student's understanding after reading a text. The third class of educational applications generates

questions to be employed in tutorial dialogues in a Socratic manner. Olney and colleagues (Olney et al., 2012) presented a method for generating questions for tutorial dialogues. This method extracts concept maps from textbooks in the domain of Biology, questions are constructed based on these concepts. Person and Graesser (2002) developed an intelligent tutoring system for the areas of computer literacy and Newtonian physics. Each topic contains a focal question, a set of good answers, and a set of anticipated bad answers. For the domain of Computer Science, Lane & VanLehn (2005) developed a tutor which is intended to help students develop pseudo-code solution to a given problem.

In the contrast to traditional approaches to generating questions using texts as an input, Jouault and Seta (2013) proposed to generate questions by querying semantic information from Wikipedia to facilitate learners' self-directed learning. Using this system, students in self-directed learning are asked to build a timeline of events of a history period with causal relationships between these events given an initial document. The student develops a concept map containing a chronology by selecting concepts and relationships between concepts from the given initial Wikipedia document to deepen their understandings. While the student creates a concept map, the system integrates the concept to its map and generates its own concept map by referring to semantic information of Wikipedia. The system's concept map is updated with every modification of the student and enriched with related concepts that can be extracted from Wikipedia. Thus, the system's concept map always contains more concepts than the student's map. Using these related concepts and their relationships, the system generates questions for the student to lead to a deeper understanding without forcing to follow a fixed path of learning.

We propose to use WordNet as a semantic source for generating questions that aim at stimulating the brainstorming of students during the process of argumentation. The approach to be presented is different from the work of Jouault and Seta in that we use natural language techniques to extract key concepts that serve as inputs to query semantic information from WordNet whereas Jouault and Seta focused on exploiting linked data technologies to extract semantic information.

## 3. Question Generation Approach

In this section, we describe conceptually how questions can be generated. A detailed description of this approach is referred to Le et al. (2014b). In order to illustrate the question generation approach, we will use the following example discussion topic that can be given to students in a discussion session:

"*The catastrophe at the Fukushima power plant in Japan has shocked the world. After this accident, the Japanese and German governments announced that they are going to stop producing nuclear energy. Should we stop producing nuclear energy and develop renewable energy instead?*"

The question generation approach consists of four steps: 1) analyzing a text structure and identifying key concepts, 2) generating questions using key concepts in a discussion topic, 3) generating questions using related concepts in WordNet, and 4) generating questions using example sentences in WordNet.

### *3.1 Analyzing text structure and identifying key concepts*

In order to automatically recognize key concepts of a discussion topic, a natural language parser is used to analyze the grammatical structure of a sentence for its constituents, resulting in a parse tree showing their syntactic relation to each other. The language parser analyzes a text and identifies the category of each constituent, for instance: determiner, noun, or verb. Since nouns and noun phrases can be used as key concepts in a discussion topic, we select from the parsing results of a discussion topic the constituents which are tagged as nouns (NN) or noun phrases (NP). In our example discussion topic from above, the following noun phrases can serve as key concepts to generate questions: *catastrophe, Fukushima power plant, nuclear energy, renewable energy*.

### *3.2 Question Generation Using Key Concepts in a Discussion Topic*

The extracted key concepts are helpful for generating questions. Yet, an issue that needs to be addressed next is to determine the types of questions to be generated. Several question taxonomies have been proposed by researchers in the area question generation. Among the existing question taxonomies, the

question taxonomy for tutoring proposed by Graesser and Person (1994) has been widely used. This taxonomy consists of 16 question categories: verification, disjunctive, concept completion, example, feature specification, quantification, definition, comparison, interpretation, causal antecedent, causal consequence, goal orientation, instrumental/procedural, enablement, expectation, and judgmental. The first 4 categories were classified as simple/shallow, 5-8 as intermediate and 9-16 as complex/deep questions. We apply this question taxonomy to define appropriate question templates for generating questions. For example, Table 1 defines some question templates for the classes "Definition" and "Feature/Property", where X is a placeholder for a key concept extracted from a discussion topic. For example, the question templates of the class "Definition" can be filled with the noun phrase "nuclear energy" and result in the following questions: *What is **nuclear energy**? What do you have in mind when you think about **nuclear energy**? What does **nuclear energy** remind you of?*

Table 1: Question Templates proposed for question generation.

| **Type** | **Question** |
|---|---|
| Definition | What is <X>? |
| | What do you have in mind when you think about <X>? |
| | What does <X> remind you of? |
| Feature/Property | What are the properties of <X>? |
| | What are the (opposite)-problems of <X>? |

### *3.3 Question Generation Using Related Concepts in WordNet*

In order to generate questions that are related to key concepts of a discussion (but which do not literally contain these concepts), sources of semantic information can be exploited (e.g., Wikipedia, Wiktionary, or WordNet). Currently, we deploy WordNet (Miller, 1995) because it is suitable to find related concepts for a discussion topic. WordNet is an online lexical reference system for English. Each noun, verb, or adjective represents a lexical concept and has a relation link to hyponyms which represent related concepts. In addition, for most words WordNet provides example sentences which can be used for generating questions. For example, if we input the word "energy" into WordNet, an example sentence like "energy can take a wide variety of forms" for this word is available. If we look for some hyponyms for this word, WordNet provides a list of direct hyponyms and a list of full hyponyms. The list of direct hyponyms provides concepts which are directly related to the word being searched. For example, the direct hyponyms of "energy" as listed by WordNet include "activation energy", "alternative energy", "atomic energy", "binding energy", "chemical energy", and more. The list of full hyponyms contains a hierarchy of hyponyms which represent direct and indirect related concepts of the word being searched. One of the advantages of WordNet is that it provides accurate information (e.g., hyponyms) and grammatically correct example sentences. For this reason, we exploit hyponyms provided by WordNet to generate questions which are relevant and related to a discussion topic. Placeholders in question templates (cf. Table 1) can be filled with appropriate hyponym values for generating questions. For example, the noun "energy" exists in the discussion topic, so that WordNet suggests "activation energy" as a hyponym. The question templates of the class "Definition" can then be used to generate questions such as: *What is **activation energy**? What do you have in mind when you think about **activation energy?** What does **activation energy** remind you of?*

### *3.4 Question Generation Using Example Sentences in WordNet*

In addition to using hyponyms, we propose to make use of example sentences to generate questions. There are existing tools which convert texts into questions. For example, ARK [13] is a syntax-based tool for generating questions from English sentences or phrases. The system operates on syntactic tree structures and defines transformation rules to generate questions. For example, a direct hyponym of the key concept "catastrophe" is "tsunami" for which there is an example sentence "a colossal tsunami destroyed the Minoan civilization in minutes". Using ARK, the example sentence can be converted into questions: "*What destroyed the Minoan civilization in minutes?*", "*When did a colossal tsunami destroy the Minoan civilization?*", "*What did a colossal tsunami destroy in minutes?*"

## 4. Evaluation

The goal of our evaluation is to determine whether automatically generated questions are perceived as as natural as human generated questions. Our study design is similar to the Turing test that requires humans to decide whether they are interacting with an actual computer program or with a human via computer mediation. The study being presented in this paper is a variation of the Turing test: we wanted to know whether automatically generated questions can be distinguished from human generated questions easily by human raters. Human raters we employed in this study were Computer Scientists (including Professors, Senior Researchers, and Phd. candidates).

### *4.1 Design*

In the first evaluation phase, we asked eight human experts from the research communities of computer based argumentation and question generation research to generate questions for three discussion topics. We received 54 questions for Topic 1, 47 questions for Topic 2, and 40 questions for Topic 3. These questions are referred to as human generated questions in this paper.

**Topic 1:** *The catastrophe at the Fukushima power plant in Japan has shocked the world. After this accident, the Japanese and German governments announced that they are going to stop producing nuclear energy. Should we stop producing nuclear energy and develop renewable energy instead?*
**Topic2:** *Recently, although the International Monetary Fund announced that growth in most advanced and emerging economies was accelerating as expected. Nevertheless, deflation fears occur and increase in Europe and the US. Should we have fear of deflation?*
**Topic 3:** *"In recent years, the European Central Bank (ECB) responded to Europe's debt crisis by flooding banks with cheap money…ECB President has reduced the main interest rate to its lowest level in history, taking it from 0.5 to 0.25 percent" . How should we invest our money?*

For each discussion topic, the system generated several hundred questions (e.g., 844 questions for Topic 1), because for each discussion topic several key concepts are extracted, and each key concept has a set of hyponyms that are queried from WordNet. For each key concept and each hyponym, fourteen questions have been generated based on defined question templates (see examples in Table 1). Since this set of generated questions was too big for a human expert evaluation, we selected a small subset of these questions manually so that the proportion between the automatic generated questions and the human generated questions was about 1:3. There were two reasons for this proportion. First, if there had been too many automatically generated questions, this could have influenced the "overall picture" of human generated questions. Second, we needed to make a trade-off between having enough (both human-generated and system-generated) questions for evaluation and considering a moderate workload for human raters. The numbers of automatically generated questions and of human generated questions are shown in Table 2.

Then, we mixed human generated questions with automatic generated questions and asked human raters to decide for each question from the mixed set whether they believed it was generated by a computer system or by a human expert. Note that these human raters were not the same human experts who generated the questions and did not know about the proportion between human-generated and system-generated questions. Specifically, the following question was answered by human raters: *Is that an automatic system-generated question? (Yes/No)*

Table 2: Number of questions generated by human experts and by the system for evaluation.

| | **Topic 1: No. of questions** | **Topic2: No. of questions** | **Topic 3: No. of questions** |
|---|---|---|---|
| Human-generated | 54 | 47 | 40 |
| System-generated | 16 | 15 | 13 |
| Total | 70 | 62 | 53 |

*4.2 Results*

We use the balanced F-score to evaluate the ratings of humans. This score is calculated based on precision and recall using the following formula:

$$F = \frac{2 * precision * recall}{precision + recall}$$

The precision for a class is the number of true positives (i.e., the number of system-generated questions correctly labeled as system-generated) divided by the total number of elements labeled as positive (i.e., labeled as system-generated), while the recall for a class is the number of true positives divided by the total number of elements that actually are positive (i.e., that are system-generated). If the F-score is high (close to 1), it shows that the system-generated questions are easy to distinguish from human-generated questions, and vice versa.

Table 3: Classification result of two raters on authorship of questions.

| | SGQ predicted by Rater 1 (% of total) | HGQ predicted by Rater 1 (% of total) | SGQ predicted by Rater 2 (% of total) | HGQ predicted by Rater 2 (% of total) | Total |
|---|---|---|---|---|---|
| **Topic 1** | | | | | |
| System-GQ | 12 (75%) | 4 (25%) | 13 (81%) | 3 (19%) | 16 |
| Human-GQ | 45 (83%) | 9 (17%) | 22 (41%) | 32 (59%) | 54 |
| **Topic 2** | | | | | |
| System-GQ | 13 (87%) | 2 (13%) | 15 (100%) | 0 | 15 |
| Human-GQ | 24 (51%) | 23 (49%) | 28 (60%) | 19 (40%) | 47 |
| **Topic 3** | | | | | |
| System-GQ | 10 (77%) | 3 (23%) | 12 (92%) | 1 (8%) | 13 |
| Human-GQ | 27 (67%) | 13 (33%) | 30 (75%) | 10 (25%) | 40 |

Table 3 shows how two human raters rated the mixed set of questions in the context of Topic 1. A high number (75%) of system-generated questions (SGQ) and 17% of human-generated questions (HGQ) have been correctly identified by this rater, resulting in a low F-score of 0.329 (Recall=0.211, Precision=0.75) that indicates that it was difficult for the rater to identify system-generated questions. This is because the first rater decided wrongly on 83% of the human-generated questions. The second rater, however, achieved a medium F-score of value 0.51 (Recall=0.371, Precision=0.813) that is higher than of the first rater, indicating that also this rater had some difficulties in distinguishing between human-generated and system-generated questions. Interestingly, although both raters had difficulties in distinguishing between human-generated and system-generated questions, the agreement between the two was poor in addition (Kappa=0.086).

In the context of Topic 2, the first rater achieved an F-score of 0.5 (Recall=0.351, Precision=0.867) The second rater showed a similar tendency with an F-score of 0.517 (Recall=0.349, Precision=1). The Kappa value for their agreement was 0.233, which can be considered as fair.

In the context of Topic 3, one question ("What is cheap money?") was generated by a human expert and by the system in identical form. This was left out from analysis (however, this question was classified as a system-generated question by both human raters). Specifically, the first rater achieved a low F-score of 0.4 (Recall=0.27, Precision=0.769). This can be explained by the fact that the first rater classified 67% of human-generated questions as generated by the system. The second rater achieved a similarly low F-score of 0.436 (Recall=0.286, Precision=0.923). Similar to the case of Topic 2, the agreement between the first and the second raters in the context of Topic 3 was fair (Kappa=0.263).

In summary, we have learned that for all raters and all three topics it was difficult to identify system-generated questions within the set of mixed questions (F-scores between 0.329 and 0.517). This is an indication that the system-generated questions appeared as natural as the human-generated questions to these raters. The agreement between raters was poor or fair, further strengthening this argument (there was little agreement on questions that seemed clearly human-generated or clearly system-generated).

## 5. Conclusions, Discussion and Future Work

In this paper, we have presented an approach to generating questions using WordNet as a source of semantic information. The goal is using generated questions to stimulate students brainstorming and thus, participate more actively in argumentation. We have conducted a pilot study comparing system-generated questions with questions that have been generated manually by researchers of the argumentation and question generation research communities. The study results show that the difference between human-generated and system-generated questions is not large: human raters could not tell the difference easily. However, it needs to be noted that we had to select manually a small number of questions from a huge amount (several hundreds) of system-generated questions. At present, we do not use an automatic algorithm for this task. We also think about limiting the number of system-generated questions, because if a student requests questions for developing arguments and s/he receives such a huge amount of questions, this can impact on the argumentation process negatively. As future work, we will develop criteria to limit the number of system-generated questions and evaluate the system-generated questions with respect to the quality and usefulness.

## References


Berthold, K., Röder, H., Knörzer, D., Kessler, W., Renkl, A. (2011). The double-edged effects of explanation prompts. *Computers in Human Behavior 27(1)*, 69–75.

Brown, J., Frishkoff, G., & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. *Proceedings of Human Language Technology Conference and Empirical Methods in NLP*, 819-826.

Graesser, A. C. & Person, N. K. (1994). Question Asking during Tutoring. *American Educational Research Journal*, 31(1), 104 –137.

Graesser, A. C., Rus, V., D'Mello, S. K., & Jackson, G. T. (2008). AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In: Robinson & Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning*, 95-125, Information Age Publishing.

Heilman, M. & Smith, N. A. (2009). Question generation via over-generating transformations and ranking. *Report CMU-LTI-09-013*, Language Technologies Institute, School of Computer Science, CMU.

Jouault, C., & Seta, K. (2013). Building a Semantic Open Learning Space with Adaptive Question Generation Support. *Proceedings of the 21st International Conference on Computers in Education*, 41-50.

Kunichika, H., Katayama, T., Hirashima, T. & Takeuchi, A. (2001). Automated Question Generation Methods for Intelligent English Learning Systems and its Evaluation. *Proceedings of the International Conference on Computers in Education*, 1117-1124.

Lane, H. C. & Vanlehn, K. (2005). Teaching the tacit knowledge of programming to novices with natural language tutoring. *Journal Computer Science Education*, 15, 183–201.

Le, N.-T., Kojiri, T. & Pinkwart, N. (2014). Automatic Question Generation for Educational Applications – The State of Art. *Advanced Computational Methods for Knowledge Engineering*, Vol. 282, 325-338.

Le, N.-T., Nguyen, N.-P., Seta, K. & Pinkwart, N. (2014). Automatic question generation for supporting argumentation. *Vietnam Journal of Computer Science* 1(2), 117-127.

Liu, M., Calvo, R.A., & Rus, V. (2012). G-Asks: An Intelligent Automatic Question Generation System for Academic Writing Support. *Dialogue and Discourse* 3 (2), 101-124.

Miller, G. A. (1995). WordNet: A lexical database. In: Communications of the ACM, 38(11), 39-41.

Mitkov, R., Ha, L. A., & Karamanis, N. (2006) A computer-aided environment for generating multiple-choice test items. *Journal Natural Language Engineering* 12 (2): 177–194. Cambridge University Press.

Mostow, J. & Chen, W. (2009). Generating instruction automatically for the reading strategy of self-questioning. *Proceeding of the Conference on Artificial Intelligence in Education*, 465-472.

Olney, A.M., Graesser, A., & Person, N.K. (2012) Question Generation from Concept Maps. *Dialogue and Discourse* 3 (2), 75–99.

Person, N. K., & Graesser, A. C. (2002). Human or Computer? AutoTutor in a Bystander Turing Test. *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, Stefano A. Cerri, Guy Gouardères, and Fábio Paraguaçu (Eds.), Springer-Verlag, 821-830.

Tenenberg, J. & Murphy, L. (2005). Knowing What I Know: An Investigation of Undergraduate Knowledge and Self-Knowledge of Data Structures. *Journal Computer Science Education*, 15(4), 297-315.

Turing, A. M. (1950). Computing machinery and intelligence. In E. A. Feigenbaum & J. Feldman (Eds.) Computers and thought. New York: McGraw-Hill.