Using Semantic Web for Generating Questions: Do Different Populations Perceive Questions Differently?

Nguyen-Thinh Le

Humboldt-Universität zu Berlin
Department of Informatics
Research Group "Computer Science Education / Computer Science and Society"
Unter den Linden 6, 10099 Berlin, Germany

Nguyen-thinh.le@hu-berlin.de

Abstract. In this paper, I propose an approach to using semantic web data for generating questions that are intended to help people develop arguments in a discussion session. Applying this approach, a question generation system that exploits WordNet for generating questions for argumentation has been developed. This paper describes a study that investigates a research question of whether different populations perceive questions (either generated by a system or by human experts) differently. To conduct this study, I asked eight human experts of the argumentation and the question generation communities to construct questions for three discussion topics and used a question generation system for generating questions for argumentation. Then, the author invited three groups of researchers to rate the mix of questions: 1) computer scientists, 2) researchers of the argumentation and question generation communities, and 3) student teachers for Computer Science. The evaluation study showed that human-generated questions were perceived differently by three different populations over three quality criteria (the understandability, the relevance, and the usefulness). For system-generated questions, the hypothesis could only be confirmed on the criteria of relevance and usefulness of questions. This contribution of the paper motivates researchers of question generation to deploy various techniques to generate questions adaptively for different target groups.

Keywords: Semantic web, Linked Open Data, Question Generation, Question Taxonomy, Adaptivity

1 Introduction

Asking questions is an important skill that is required in many institutional settings, e.g., interviews conducted by journalists (Clayman & Heritage, 2002), medical settings (Drew & Heritage, 1992), courtrooms (Atkinson & Drew, 1979). For teachers, asking questions is almost an indispensable teaching technique. Dillon (1988) investigated questions generated by teachers in 27 upper classrooms in six secondary schools in the

USA and reported that questions accounted for over 60% of the teachers' talk. The benefits of using questions in instruction are multi-faceted and have been reported in many research studies (Lin et al., 2014; Morgan & Saxton, 2006; Tenenberg & Murphy, 2005). Not only teachers' questions can enhance learning, but also students' question asking can benefit learning. The evidence from research studies provides a solid empirical basis to support the inclusion of students' question asking in teaching in order to enhance comprehension (Rothstein & Santana, 2014), cognitive and metacognitive strategies use (Yu & Pan, 2014), and problem-solving abilities (Barlow & Cates, 2006) of students. Researchers suggested that teachers should pose questions that encourage higher-level thinking of students because they need to be familiarized with different levels of thinking and to use knowledge of the lower-level productively (Chafi & Elkhozai, 2014). In addition, Morgan and Saxton (2006) demonstrated that well-chosen higher-order questions can not only be used to assess student's knowledge but also to extend his/her knowledge, to improve his/her skills of comprehension and application of facts and also to develop his/her higher-order thinking skills. Yet the evidence is that the majority of questions teachers use in their classrooms in order to check knowledge and understanding, to recall of facts or to diagnose student's difficulties (Chafi & Elkhozai, 2014), and only about 10% of questions are used to encourage students to think (Brown & Wragg, 1993). Especially, pre-service teachers, who have just graduated their study, would have many difficulties in generating questions in their classes.

Many automatic question generation approaches have been developed in order to help teachers and students. For example, in the LISTEN[1] project (Mostow & Chen (2009), Mostow & Beck (2007)), Mostow and colleagues developed an automated reading tutor which deploys automatic question generation to improve the comprehension capabilities of students while reading a text. Kunichika et al. (2001) proposed an approach to extracting syntactic and semantic information from an original text and questions are constructed using the extracted information to support novices in learning English. Heilman and Smith (2010) developed an approach to generating questions for assessing students' acquisition of factual knowledge from reading materials. What all these approaches have in common is that they deployed information in a given text (e.g., a reading text) to generate questions.

This paper proposes to use existing encyclopedic or lexical knowledge databases available on the Internet as semantic sources for generating questions automatically. Using the semantic web as a source of information required for generating questions may save time for teachers in preparation for their lessons. Currently, this approach has been experimented by Jouault and Seta (2014) who proposed to generate semantics-based questions by querying 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 a given initial Wikipedia document to deepen their understandings. While the student creates a concept map, the system also integrates the concept to its map and generates its own concept map by

[1] http://www.cs.cmu.edu/~./listen/

referring to semantic information of Wikipedia. The system's concept map is updated with every modification of the student's one 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. Also exploiting semantic web data sources, Le et al. (2014) proposed to use WordNet to generate questions that are intended to help students develop arguments for discussion. Their project aimed at using automatically generated questions for stimulating the brainstorming of students during the process of argumentation. WordNet has been used as a semantic source for generating questions, because it is a rich lexical database that is able to provide hyponyms (related concepts) to a queried concept. In a recent study, Le and Pinkart (2015) investigated the quality of system-generated questions with respect to the understandability, the relevance of questions to a given discussion topic, and the usefulness of questions for students to develop new arguments. The authors reported that system-generated questions could not be distinguished from human-generated questions in the context of two discussion topics (topic about nuclear energy and topic about low interest rate) while the difference between system-generated questions and human-generated questions was noticed in the context of one discussion topic (deflation in Europe and US).

This paper not only investigates the quality of system-generated questions, but also the hypothesis that people of different populations perceive the quality of questions differently. That is, questions that are understandable, relevant to a discussion topic, and useful for teachers might be perceived not understandable, irrelevant to a discussion topic, not useful by students. This hypothesis will be investigated based on not only human-generated questions, but also questions that are generated by the question generation system developed by Le et al. (2014) for the domain of argumentation.

The remainder of this paper is structured as follows. The next section will review sources of semantic web data sources that can be used to generate questions. Then, in the third section, the study for investigating the formulated hypothesis will be described. In the fourth section, I will discuss on the results of the study, and in the final section, the conclusions will be summarized.

2 A Review of Semantic Web and Linked Open Data Sources

In order to help students develop new arguments for the argumentation, asking questions is one of the useful strategies. In order to create questions, semantic information which is related to a given topic is required. Different semantic sources (such as semantic web data and linked open data) can serve to create questions. Presently, many useful semantic web and linked open data sources have been developed by large communities (including non-experts and experts), e.g., Wiktionary[2], Openthesaurus[3], and GermaNet[4]

[2] http://de.wiktionary.org/wiki/Wiktionary:Hauptseite
[3] https://www.openthesaurus.de/
[4] http://www.sfs.uni-tuebingen.de/GermaNet/

for the German language; WordNet[5] and Freebase[6] for the English language; BabelNet[7] and DBPedia[8] are multilingual databases; and more.

In the following, I will review only the sources of semantic web data and linked open data for the English language that are maintained continuously and have a considerable number of datasets. The review is followed by a thorough analysis with respect to their usefulness in the context of question generation for argumentation.

The purpose of YAGO is combining information from different Wikipedia databases in multiple languages. The YAGO knowledge base is automatically constructed from Wikipedia and consists of entities, facts, and relations. Each article in Wikipedia represents an entity in the knowledge base YAGO. Two entities can stand in a relation. For example, the fact **AlbertEinstein** *hasWonPrize* **NobelPrize** has the relation *hasWonPrize* that has entities **AlbertEinstein** and **NobelPrize.** For the purpose of generating questions for helping develop new arguments, such relations may be useful. For example, we can generate a question using the relation *hasWonPrize:* "Which prize did Albert Einstein win?" This question may stimulate students to think about Einstein's work achievements for that prizes were announced. The version YAGO2 has over 9.8 million entities and 447 million facts (Hoffart et al., 2013). The YAGO3 version has 77 English relations (Mahdisoltani et al., 2015).

WordNet (Miller, 1995) also provides a source of semantic information which can be related to a discussion topic. WordNet is an online lexical reference system for English. Each noun, verb, or adjective represents a lexical concept. A concept is represented as a synonym set (called synset), i.e., the set of words that share the same meaning. Between two synsets, WordNet provides semantic relations (12 relations for nouns). The hyponym relation represents a concept specialization. For example, for the concept "energy", WordNet provides a list of direct hyponyms which are directly related to the concept being searched and represent specializations: "activation energy", "alternative energy", "atomic energy", "binding energy", "chemical energy", and more. In addition, synsets can contain sample sentences to provide sample sentences, which can be used for generating questions. For example, if we input the word "energy" into WordNet, an example sentence, e.g., "Energy can take a wide variety of forms" for this concept is available. Sample sentences provided by WordNet may also be exploited to create questions, e.g.: "Which forms can energy take?" One of the advantages of WordNet is that it provides accurate information (e.g., hyponyms) and grammatical correct sample sentences which may serve useful semantic information for generating questions.

BabelNet (Navigli & Ponzetto, 2012) is a multilingual semantic network which is an integration of lexicographic and encyclopedic knowledge from WordNet and Wikipedia. In addition to the standard WordNet relations, BabelNet is enriched with "gloss" relations and unlabeled relations that are derived from internal links in the Wikipages. A gloss relation is established based on a gloss for a concept in WordNet. For example, the gloss of the first synset of "play" is "a dramatic work intended for performance by

[5] http://wordnet.princeton.edu/
[6] https://www.freebase.com/
[7] http://babelnet.org/
[8] http://dbpedia.org

actors on a stage", and so the first sense of "play" is gloss-related with the first sense of "actor" and the third sense of stage (in WordNet, each lexical unit may have several senses). Since Wikipages typically contain hypertext linked to other Wikipages, thus, it refers to related concepts. For instance, "play" (with sense "theatre") has links to "literature", "playwright", etc. BabelNet exploits these links in order to extend the relations between concepts in its database. In the current version (Navigli & Ponzetto, 2012), BabelNet has 51,087,221 relations for the English language and this number of relations is enormously higher than the number of relations provided by WordNet (364,522). However, with respect to using gloss-relations and wikipages-relations for generating questions, I do not see benefits because these relations do not provide a specific semantic relationship between two concepts in order to generate a meaningful question. BabelNet is more useful than WordNet with regard to generating questions for different languages, because BabelNet is a multi-lingual database, while WordNet just supports the English language.

3 Do Different Populations perceive Questions Differently?

The title of this section is the research question of this paper. Either questions are developed by human experts or generated by a computer systems, it is interesting for the community of question generation to know whether people of different populations perceive them differently. This research question is important because it helps us understand more about how people perceive questions, and thus, consequently, question generators (human or system) might have to adapt questions to target persons. In order to investigate this research question, I use the question generation system that has been developed by Le at al. (2014) with the intention to support the process of argumentation. In this paper, I briefly summarize the approach to generating questions using WordNet developed by Le and colleagues (2014). The authors exploited two types of semantic information for generating questions. First, questions are generated using key concepts in a discussion topic. For example, the following discussion topic 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?*"

From the discussion topic, the system extracts nouns and noun phrases to serve as key concepts for generating questions: *catastrophe, Fukushima power plant, nuclear energy, renewable energy*. These nouns and noun phrases are filled in a set of predefined question templates, and as a result, a set of questions are generated.

Table 1 shows a part of the set of pre-defined question templates implemented in the question generation system, where the left column represents the type of questions

which can be instantiated by filling in the place-holder <X> of the corresponding template (the right column). The authors applied the question taxonomy developed by Graesser et al. (1992) and developed fourteen question templates.

Table 1. Question templates for question generation.

Type	Question template
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>?
	What features does <X> have?

The second type of information is using hyponyms provided in WordNet for each concept (cf. Section 2). Placeholders in pre-defined question templates can be filled with appropriate hyponym values for generating questions. For example, the noun "energy" exists in the discussion topic, and after extracting this noun as a key concept, it can be used as input for WordNet that provides several hyponyms, including "activation energy". The following question templates (Table 2, 2nd column) can be used to generate questions of the question type "Definition".

Table 2. Question templates and generated questions of the type „Definition".

Type	Question template	Question
Definition	What is <X>?	What is activation energy?
	What do you have in mind when you think about <X>?	What do you have in mind when you think about activation energy?
	What does <X> remind you of?	What does activation energy remind you of?

In addition to using hyponyms for generating questions, Le at al. (2014) proposed to use example sentences provided by WordNet for each concept to generate questions. For example, the following questions have been generated using the sample sentence that is provided in WordNet "*catalysts are said to reduce the energy of activation during the transition phase of a reaction*"

- *Are catalysts said to reduce the energy of activation during the transition phase of a reaction?*
- *When are catalysts said to reduce the energy of activation?*
- *What are catalysts said to reduce during the transition phase of a reaction?*
- *What are said to reduce the energy of activation during the transition phase of a reaction?*
- *What are catalysts said to reduce the energy of during the transition phase of a reaction?*

Since in a pre-study Le et al. (2014) found that most questions generated using sample sentences provided by WordNet do not represent meaningful question items, this type of semantic information (sample sentences) for generating questions is opted out of the study that will be described in the next section.

3.1 Study design

The goal of this study is to investigate how people from different populations perceive questions that are generated by human experts and by a computer system. For this purpose, the author invited three groups of human raters to join the study. The first group included seven computer scientists who are professors or PhD students of Computer Science. The second group is represented by six senior researchers of the argumentation and the question generation communities. The third group included six student teachers who are studying Computer Science Education at the Humboldt Universität zu Berlin and all of them are native Germans. They can understand English properly, because all German high school students must study English as the first foreign language.

The study consists of two phases. First, eight experts from the research communities of argumentation and question/problem generation (six of them are in the second population of human raters) were invited to manually create questions. They got the following three discussion topics by emails and were asked to create questions which can be used to support students in developing arguments. Since the eight experts work in the USA, Europe and Asia, the discussion domains were chosen with international relevance and had been in the news recently. For this study, the domains of energy and economy were chosen. Each discussion topic consisted of two sentences and an initial discussion question. The intention of this kind of construction for discussion topics was that the discussion participants and the human experts should have enough "materials" for thinking about a specific problem. If a discussion topic was too short (e.g., only a sentence or a discussion question), this might make it difficult for discussion participants to initiate a discussion or for human experts to think of questions to be generated:

> **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"[9] *. How should we invest our money?*

From eight experts, 54 questions for Topic 1, 47 questions for Topic 2, and 40 questions for Topic 3 were received.

Then, the same discussion topics were input into the question generation system for argumentation. For each discussion topic, the system generated several hundred questions (e.g., 844 questions for Topic 1), because from each discussion topic several key concepts were extracted, and each key concept was extended with a set of hyponyms queried from WordNet. For each key concept and each hyponym, fourteen questions have been generated based on fourteen pre-defined question templates. Since the set of generated questions was too big for expert evaluation, a small amount of automatic generated questions was selected randomly, 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, in case the proportion between automatically generated questions and human generated questions is too high, then it could influence the real "picture" of human generated questions. Second, a trade-off between having enough (both human-generated and system-generated) questions for evaluation and a moderate workload for human raters needs to be considered. The proportion of automatic generated questions and human generated questions is shown in Table 3.

Table 3. Number of questions generated by human experts and by the system.

	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

In the second phase of the study, the whole list of human-generated and system-generated questions was given to the first group (computer scientists) and the third group (student teachers). For each individual human rater of the second group, an individual list of mixed questions was created, i.e., the questions which have been generated by each senior researcher of the argumentation and the question generation communities were removed, because they would identify the questions created by themselves easily. Thus, the list of questions to be rated by the second group was shorter than the list of questions for the first and the third groups. All raters were asked to rate each question based on a scale between 1 (bad) and 3 (good) over three quality criteria: the understandability of a question, the relevance of a question to a given discussion topic, the usefulness of a question for students to develop new arguments.

[9] http://www.spiegel.de/international/europe/ecb-surprises-economists-by-dropping-key interest-rate-to-historic-low-a-932511.html

3.2 Results: Quality difference between system-generated and human-generated questions

In order to determine the quality difference between system-generated and human-generated questions, two-sided t-tests are used and p-values are taken as indicators for statistical significance of the difference in quality between the two groups.

Table 4. Quality of questions for Topic 1

Group 1: Computer Scientists			
	Understandability Mean (s.d.)	**Relevance Mean (s.d.)**	**Usefulness Mean (s.d.)**
System-GQ	2.19 (0.89)	1.96 (0.87)	1.69 (0.69)
Human-GQ	2.28 (0.80)	2.14 (0.86)	2.12 (0.87)
Difference	t=0.67	t=1.25	t=0.39
Significance	p=0.51 (not significant)	p=0.21 (not significant)	p=0.0009 (significant)
Group 2: Argumentation and Question Generation researchers			
System-GQ	2.53 (0.57)	1.81 (0.59)	1.5 (0.76)
Human-GQ	2.53 (0.62)	2.51 (0.60)	2.35 (0.68)
Difference	t=0.04	t=5.67	t=5.91
Significance	p=0.9682 (not significant)	p<0.0001 (significant)	p<0.0001 (significant)
Group 3: Student teachers			
System-GQ	2.13 (0.79)	1.47 (0.72)	1.44 (0.67)
Human-GQ	2.82 (0.43)	2.65 (0.55)	2.67 (0.58)
Difference	t=6.52	t=9.87	t=10.16
Significance	p<0.0001 (significant)	p<0.0001 (significant)	p<0.0001 (significant)

Table 4 shows the quality of human-generated questions and system-generated questions for Topic 1. From this table we can notice that for the group of student teachers, the human-generated questions are significant better than system-generated questions with respect to their understandability (t=6.52, p<0.0001), their relevance to the given discussion topic (t=9.87, p<0.0001), their usefulness for helping students develop new arguments (t=10.16, p<0.0001). From the perspective of computer scientists and researchers of the argumentation and question generation communities, system-generated questions deserved better ratings with respect to understandability: the difference between system-generated questions and human-generated questions was statistically not significant (Group 1: t=0.67, p=0.51; Group 2: t=0.04, p=0.9642). With respect to the relevance of questions, the second and the third groups agreed that human-generated questions are significantly better than system-generated questions (Group 2: t=5.67, p<0.0001; Group 3: t=9.87, p<0.0001). With respect to the usefulness of questions, the

ratings of three groups indicated that human-generated questions are significantly better than system-generated questions (Group 1: t=0.39, p=0.0009; t=5.91, p<0.0001; Group 3: t=10.16, p<0.0001). That is, all three groups agreed that system-generated questions are not useful as human-generated questions for helping students develop new arguments. In overall, the first and the second groups rated system-generated questions higher than average (since the rating scale is from 1 to 3, the averaged value is 1.5) whereas the group of student teachers only rated the understandability of system-generated questions higher than average (m=2.13, s.d.=0.79).

Table 5. Quality of questions for Topic 2

	Understandability Mean (s.d.)	**Relevance Mean (s.d.)**	**Usefulness Mean (s.d.)**
Group 1: Computer Scientists			
System-GQ	2.40 (0.77)	1.83 (0.7)	1.87 (0.82)
Human-GQ	2.76 (0.48)	2.43 (0.73)	2.37 (0.70)
Difference	t=3.01	t=3.93	t=3.29
Significance	p=0.0031 (significant)	p=0.0001 (significant)	p=0.0013 (significant)
Group 2: Argumentation and Question Generation researchers			
System-GQ	1.97 (0.89)	1.60 (0.77)	1.50 (0.68)
Human-GQ	2.61 (0.64)	2.56 (0.67)	2.39 (0.72)
Difference	t=4.21	t=6.46	t=5.90
Significance	p=0.0001 (significant)	p=0.0001 (significant)	p=0.0001 (significant)
Group 3: Student teachers			
System-GQ	2.13 (0.86)	1.73 (0.58)	1.37 (0.56)
Human-GQ	2.74 (0.60)	2.47 (0.63)	2.34 (0.71)
Difference	t=4.33	t=5.63	t=6.85
Significance	p=0.0001 (significant)	p=0.0001 (significant)	p=0.0001 (significant)

Table 5 shows the quality of questions that have been generated for Topic 2 by human experts and by the system. The first point we can learn from results shown in this table is that all groups of human raters rated human-generated questions significantly better than system-generated questions. This result for Topic 2 is consistent with the conclusion of the evaluation study conducted by Le and Pinkwart (2015). The second point is that system-generated questions have been rated over average (e.g., Group 1 rated 2.4 for understandability, Group 2 rated 1.6 for the relevance criterion, Group 3 rated 1.73 for the relevance criterion) by three groups except the rating for the usefulness criterion given by the group of student teachers (1.37). The similar picture for human-generated questions can also be identified: with respect understandability, three groups rated between 2.61 and 2.76; for the relevance criterion, the ratings are between

2.43 and 2.56; and for the usefulness criterion the ratings given each group is almost the same (between 2.34 and 2.39).

Table 6. Quality of questions for Topic 3

	Understandability Mean (s.d.)	**Relevance Mean (s.d.)**	**Usefulness Mean (s.d.)**
Group 1: Computer Scientists			
System-GQ	2.21 (0.72)	1.92 (0.88)	1.71 (0.69)
Human-GQ	2.27 (0.80)	2.21 (0.78)	2.04 (0.76)
Difference	t=0.33	t=1.54	t=1.89
Significance	p=0.7397 (not significant)	p=0.1272 (not significant)	p=0.0613 (not significant)
Group 2: Argumentation and Question Generation researchers			
System-GQ	2.25 (0.85)	1.92 (0.88)	1.50 (0.83)
Human-GQ	2.50 (0.74)	2.50 (0.68)	2.24 (0.85)
Difference	t = 1.37	t = 3.34	t = 3.67
Significance	p=0.1754 (not significant)	p=0.0012 (significant)	p=0.0004 (significant)
Group 3: Student teachers			
System-GQ	2.38 (0.88)	2.17 (0.76)	2.25 (0.74)
Human-GQ	2.76 (0.51)	2.59 (0.59)	2.47 (0.68)
Difference	t = 2.65	t = 2.70	t = 1.39
Significance	p=0.0093 (significant)	p=0.0081 (significant)	p=0.1682 (not significant)

Table 6 shows the quality of generated questions for Topic 3. First, we can see that the given ratings are not consistent among three groups of human raters. The ratings of the group computer scientists for system-generated questions with respect to the understandability (t=0.33, p=0.7397), the relevance (t=1.54, p=0.1272), and the usefulness (t=1.89, p=0.0613) are not significantly different from the ratings for human-generated questions. That means that system-generated questions are of similar quality as human-generated questions. On the contrary, the group of argumentation and question generation researchers held the human-generated questions significantly better than system-generated questions with respect to the relevance (t=3.34, p=0.0012), the usefulness (t=3.67, p=0.0004). Similarly, the group of student teachers rated the human-generated questions significantly better than the system-generated questions with respect to the understandability (t=2.65, p=0.0093) and the relevance (t=2.70, p=0.0081).

Second, it is surprising that the ratings given by the group of student teachers for system-generated questions (understandability: 2.38, relevance: 2.17, usefulness: 2.25) are higher than the ratings given by the group Computer Scientists (understandability: 2.21, relevance: 1.92, usefulness: 1.71) over three criteria, while for Topic 1 and Topic 2, the group of student teachers rated system-generated questions always lower than

human-generated questions. Why the phenomenon of disagreement on ratings between the three groups appeared and why the group of student teachers rated system-generated questions for Topic 3 better than for other discussion topics, these need further investigation.

3.3 Difference of perceiving the quality of questions between three populations of human raters

In this section, I investigate whether the quality of (both system-generated and human-generated) questions is perceived differently between three populations of human raters (computer scientists, researchers in argumentation and question generation communities, and student teachers). For this purpose, ANOVA will be used to analyze the difference of the quality of questions over three groups. ANOVA variance analysis will be performed over three independent samples. Each sample represents the ratings collected from each group of human raters. The samples are independent because for the group of argumentation and question generation researchers their own questions have been removed from the set of mixed questions to be rated. They should not rate the questions that have been generated by themselves. The other groups (computer scientists and student teachers) were assigned with a complete set of mixed questions.

Table 7 shows the difference of quality of system-generated questions (3^{rd}-4^{th} rows) and human-generated questions (6^{th}-7^{th} rows) perceived by three groups of human raters. These questions have been generated for the discussion topic about stopping nuclear energy (Topic 1). With respect to the understandability and the usefulness, while the difference in quality of system-generated questions between three groups of human raters is statistically not significant (understandability: $p=0.08$, usefulness: $p=0.26$), the difference in quality of human-generated questions between three groups of human raters is statistically significant ($p<0.0001$ over all three criteria). This indicates that the three groups of human raters perceived the quality of human-generated questions for Topic 1 differently. Similarly, three groups of human raters rated the relevance of system-generated questions significant differently.

Table 7. Difference between three groups of human raters with respect to questions for Topic 1

	Understandability Mean (s.d.)	**Relevance Mean (s.d.)**	**Usefulness Mean (s.d.)**
System-generated questions			
Difference	F=2.6	F=4.05	F=1.38
Significance	p=0.0789 (not significant)	p=0.0201 (significant)	p=0.2559 (not significant)
Human-generated questions			
Difference	F=21.97	F=18.46	F=18.17
Significance	p<0.0001 (significant)	p<0.0001 (significant)	p<0.0001 (significant)

Table 8 shows results of ANOVA analysis over three groups of human raters for the quality of system-generated and human-generated questions. These questions have been generated for the discussion topic about fear of deflation in Europe and the USA (Topic 2). For system-generated questions, with respect to the understandability and the relevance, there is no difference in ratings between three groups (understandability p=0.1388, relevance: p=0.4226). This indicates that all three groups of human raters agreed on the high understandability of system-generated questions (m=1.97-2.40, cf. Table 5) and the moderate relevance (m=1.60-1.83, cf. Table 5). However, with respect to the usefulness, there is a significant difference between three groups of human raters (p=0.0187). This indicates that different groups of human raters perceived the usefulness of system-generated questions differently.

For human-generated questions, the ratings of human raters are consistent over all three criteria, i.e., there is no significant difference in ratings among three groups of human raters (6th-7th rows of Table 8). That is, they agreed on the high understandability (m=2.61-2.76), high relevance (m=2.43-2.56), and high usefulness (m=2.29-2.39) (cf. Table 5).

Table 8. Difference between three groups of human raters with respect to questions for Topic 2

	Understandability Mean (s.d.)	**Relevance Mean (s.d.)**	**Usefulness Mean (s.d.)**
System-generated questions			
Difference	F=2.02	F=0.87	F=4.17
Significance	p=0.1388 (not significant)	p=0.4226 (not significant)	p=0.0187 (significant)
Human-generated questions			
Difference	F=1.7	F=0.9	F=0.11
Significance	p=0.1847 (not significant)	p=0.4078 (not significant)	p=0.8959 (not significant)

Table 9 shows the difference in ratings between three groups of human raters for system-generated and human-generated questions. These questions have been developed for the discussion topic about the low interest rate in Europe (Topic 3). The table shows that with respect to the understandability and the relevance, there was no significant difference in ratings for system-generated questions between three groups of human raters (Table 9, 4th row: understandability: p=0.7642, relevance: p=0.5001). This indicates that all human raters agreed on the high understandability (m=2.21-2.38, cf. Table 6) and high relevance (m=1.92-2.17, cf. Table 6) of system-generated questions. However, with respect to the usefulness of system-generated questions, the difference in ratings between three groups is significant. This indicates that different groups of human raters perceived the usefulness of system-generated questions differently. We notice that this is also the case for Topic 2.

For human-generated questions, the difference in ratings is significantly different over all three criteria (cf. 6th-7th rows of Table 9). This indicates that three populations perceived the quality of human-generated questions, which have been developed for

Topic 3, differently, although the understandability (m=2.27-2.76, cf. Table 6), the relevance (m=2.21-2.59, cf. Table 6), and the usefulness (m=2.04-2.47, cf. Table 6) of those questions were rated highly. This phenomenon is similar to the human-generated questions for discussion Topic 1 (cf. Table 7).

Table 9. Difference between three groups of human raters with respect to questions for Topic 3

	Understandability Mean (s.d.)	**Relevance Mean (s.d.)**	**Usefulness Mean (s.d.)**
System-generated questions			
Difference	F=0.27	F=0.7	F=6.29
Significance	p=0.7642 (not significant)	p=0.5001 (not significant)	p=0.0031 (significant)
Human-generated questions			
Difference	F=9.7	F=6.41	F=6.25
Significance	p<0.0001 (significant)	p=0.0020 (significant)	p=0.0023 (significant)

In summary (cf. Table 10), three groups of human raters perceived human-generated questions differently in the context of Topics 1 and 3 over three criteria. System-generated questions, they were perceived differently by three groups of human raters with respect to specific criteria: the relevance of questions for Topic 1 and the usefulness of questions for Topics 2 and 3.

Table 10. Summary of differences between three groups of human raters

	Understandability	**Relevance**	**Usefulness**
Topic 1			
System-GQ	No	Yes	No
Human-GQ	Yes	Yes	Yes
Topic 2			
System-GQ	No	No	Yes
Human-GQ	No	No	No
Topic 3			
System-GQ	No	No	Yes
Human-GQ	Yes	Yes	Yes

4 Related Work and Discussion

Several applications of question generation for argumentation have been devised. Liu and colleagues (Liu et al., 2012) introduced a system (G-Asks) for improving students' writing skills (e.g., citing sources to support arguments, presenting the evidence in a persuasive manner). The approach implemented in this system consists of three stages. First, citations in an essay written by the student are extracted, parsed and simplified. Then, in the second stage, the citation category (opinion, result, aim of study, system, method, and application) is identified for each citation candidate. In the final stage, an appropriate question is generated using pre-defined question templates. Evaluation studies have shown that the system could generate questions as useful as human supervisors and significantly outperformed human peers and generic questions in most quality measures after filtering out questions with grammatical and semantic errors (Liu et al., 2012).

While the work of Liu and colleagues (Liu et al., 2012) focused on improving the writing skills of students, Adamson and colleagues (2013) proposed to generate discussion questions automatically in order to support instruction. Adamson and colleagues investigated three different approaches of selecting sentences from a summary text: the cosine similarity (Huang, 2008), LSA content scores (Dumais, 2004), and TF-IDF uniqueness (Wu et al., 2008). Selected representative sentences are transformed into discussion questions. In order to rank the generated questions on the basis of abstraction and ability to trigger discussion, the authors calculated the subjectivity score by averaging the subjectivity values of each word in the sentence using SentiWordNet (Baccianella et al., 2010). SentiWordNet is a database of word-senses that provides subjectivity scores assigned to each word. Discussion questions have been generated applying these three approaches and evaluated by asking four teachers for rating. With respect to stimulating discussion, the LSA and Cosine similarity approaches were significantly better than TF-IDF. In addition, the evaluation study showed that there were no significant distinctions between the approaches on the dimension of comprehensibility (i.e., a generated question is comprehensible) and important themes (i.e., a generated question touches upon important themes from the story).

Similarly to the work of Adam and colleagues (2013), the approach of generating questions I present in this paper aims at stimulating the process of developing new arguments for a discussion topic. The difference between this approach and the approaches of Liu et al. (2012) and Adamson et al. (2013) is that the question generation approach being presented in this paper deploys WordNet as a source of semantic information for generating questions.

From the results of the study in the previous section, two lessons have been learned. First, in the context of a specific topic (e.g., Topic 3), system-generated questions have similar quality as human-generated questions over all three criteria (understandability, relevance, and usefulness). Second, different groups of human raters perceive, especially, human-generated questions differently. System-generated questions have also been perceived differently, but with respect to their relevance or their usefulness.

These results are important for instructors and automatic question generators to adapt questions to individual target groups (e.g., primary school students, high school students, or university/college students). For instructors, if their questions are not understandable for students, they may explain or reformulate them in another way. For automatic question generators, it is difficult to reformulate questions. I propose two strategies for automatic question generators. First, we can define different sets of question templates for different levels of target groups. The formulation of these questions templates depends on the level of the target group. If the target group consists of intellectual university students, then question templates can be formulated in a scientific manner. Second, we can adopt different question taxonomies for different target groups. Most school teachers know the Bloom's taxonomy (Bloom, 1956), that has six levels: 1) knowledge, 2) comprehension, 3) application, 4) analysis, 5) synthesis, 6) evaluation. Teachers usually apply this question taxonomy in schools (Arias de Sanchez, 2013). Beside the Bloom's taxonomy, there are many other question taxonomies, e.g. PREG (Otero & Graesser, 2001), Schreiber (1967), Pate & Bremer (1967), among which the taxonomy developed by Graesser et al. (1992) is widely used for tutoring. To my best knowledge, until now, there is no study that compares the applicability of different question taxonomies for different target groups. However, question taxonomies may be used to individualize questions for automatic question generator.

During the second phase of the evaluation study (cf. Section 3.1), some student teachers optionally informed me about the different criteria they used to identify system-generated questions. Their criteria are very various: 1) a system-generated question is superficial with regard to a given discussion topic, 2) a system-generated question is similar to another one in the mixed set of questions, 3) a system-generated question that expects a factual answer and is intuitive (e.g., "What features does ECB president have?"), 4) a system-generated question that contains unknown information (e.g., "How will those policies affect those outcomes/stakeholders?"). To identify human-generated questions, they applied the criterion: human-generated questions may have typo/syntax errors, while system-generated questions are error-free. However, some of these criteria for guessing system-generated may be not justified, because for instance the developer of question templates could also have made typo/syntax errors as well.

Some of the human experts in the argumentation and question generation communities had more systematic criteria when they were asked to generate questions. Since I did not request all human experts to explicitly explain how they generated questions, I did not receive their strategies of generating questions. However, some of them optionally described their strategies. One of them applied different types of arguments/argumentative schemes based on the argumentum model of Rigotti and Greco Morasso (2010). In the following, I list all the argument types that have been used to generate questions to stimulate the argumentative reflection of students by one of the human experts in the argumentation community (the questions in the brackets were generated by her):

- Arguments of consequence or warning: e.g., "Think about the consequences of the nuclear catastrophe: which were (and still are) the consequences on people, in particular on their health condition?"

- Arguments of alternatives: e.g., “Think about different sources of renewable energy: which alternative sources of energy do we have/know?”
- Arguments of likeliness or difference: e.g., “Try to compare nuclear energy with sources of renewable energy: which are the differences in terms of productivity?”
- Arguments of termination and setting up: e.g., “Ponder about the possibility to go on producing nuclear energy: is it possible to make nuclear energy production safer than it is now?”
- Arguments of definition and ontological implications: “What does “deflation” mean? Is deflation bad for the economy of a country?”
- Argument of expert opinion/of authority: “Think about the authoritativeness of the information source: what is the International Monetary Fund? What does it do?”
- Argument of analogy: “Refer to similar past events: what did happen to the economies which had to face deflation?”
- Argument of final cause: “Think about the expected results of a money investment: what does an investor expect from investing money?”

Another human expert in the argumentation community explained how he generated questions. He applied several general questions when asking about a policy topic: “What stakeholders will be affected by this policy? What outcomes are the policy makers attempting to address? What other outcomes might be affected? How will this policy affect those outcomes/stakeholders? What is the evidence that this policy would affect these outcomes? What other policies are available (possible)? What are the pros/cons of each policy? Which policies have the best set of tradeoffs? For which stakeholders?” For example, for Topic 1 (nuclear energy), he generated the following questions: “What stakeholders will affected by stopping this production? What outcomes are the policy makers attempting to address? What other outcomes might be affected? How will stopping production affect those outcomes/stakeholders? What other policies are available (possible)? What are the pros/cons of each policy? Which policies have the best set of tradeoffs? For which stakeholders?” We notice that the last three questions are not related to a specific topic (e.g., nuclear energy) and can be used for general policy topics.

The argumentum model and the general questions for polity topics that were applied by two researchers of the argumentation community may be a good basis to define question templates for question generation systems that aim at generating questions to support argumentation. Whether the argumentum model and the general questions can be applied for different topics and results in meaningful questions, this needs further investigation.

5 Conclusions

In this paper, I have investigated the hypothesis that human-generated questions and computer-generated questions are perceived differently by different populations. For this purpose, a study in which eight human experts in the argumentation and question

generation communities have been invited to construct questions and a question generation system has been deployed to generate questions. In total, 141 human-generated and 44 system-generated questions have been mixed and rated by three groups of human raters: computer scientists, argumentation and question generation researchers, and student teachers. The study confirmed the hypothesis for human-generated questions over three quality criteria (understandability, relevance, and usefulness) that different populations perceive questions differently. For system-generated questions, the hypothesis could only be confirmed on the criteria of relevance and usefulness of questions. The results of this study are the contribution of this paper that propose researchers on question generation to adopt different strategies (e.g., different question taxonomies, different sets of question templates) for different target groups of questions.

6 Acknowledgements

The author would like to thank researchers of the argumentation community and the problem/question generation community (Prof. Kevin Ashley, Prof. Kazuhisa Seta, Prof. Tsukasa Hirashima, Prof. Matthew Easterday, Prof. Reuma De Groot, Prof. Fu-Yun Yu, Dr. Bruce McLaren, Dr. Silvia De Ascaniis) for generating questions, Computer Scientists (Prof. Ngoc-Thanh Nguyen, Prof. Viet-Tien Do, Dr. Thanh-Binh Nguyen, Zhilin Zheng, Madiah Ahmad, Sebastian Groß, Sven Strickroth), and student teachers at the Humboldt-Universität zu Berlin for their contribution in this evaluation study. Especially, the author would like to express his gratitude to Prof. Pinkwart for introducing experts of the argumentation community.

7 References

1. Alkinson, J. M. & Drew, P. (1979). *Order in Court*. London: Macmillan.
2. Adamson, D., Bhartiya, D., Gujral, B., Kedia, R., Singh, A., & Rosé, C. P. (2013). Automatically Generating Discussion Questions. *Artificial Intelligence in Education*, pp.81-90.
3. Arias de Sanchez, G. (2013). The Art of Questioning: Using Bloom's Taxonomy in the Elementary School Classroom. *Teaching Innovations Projects*: 3(1), Article 8.
4. Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.
5. Barlow, A. & Cates, J.M. (2006). The impact of problem posing on elementary teachers' beliefs about mathematics and mathematics teaching. *School Science and Mathematics*, 106(2), pp. 64-73.
6. Bloom, B. S. (1956): *Taxonomy of educational objectives: Handbook 1: Cognitive domain*. Addison Wesley Publishing
7. Brown, G. & Wragg, E. C. (1993). *Questioning.* Routledge.
8. Chafi, M.E. & Elkhouzai, E. (2014). Classroom Interaction: Investigating the Forms and Functions of Teacher Questions in Moroccan Primary School. *Journal of Innovation and Applied Studies*, 6(3), pp. 352-361.
9. Clayman, S. & Heritage, J. (2002). *The News Interview: Journalists and Public Figures on the Air*. New York: Cambridge University Press.

10. Dillon, J. T. (1988). Questioning and Teaching. *A Manual of Practice*. Croom Helm.
11. Drew, P & Heritage, J (1992). Analyzing talk at work: an Introduction. In Drew, P. & Heritage, J. (eds.), *Talk at Work: Interaction in Institutional Settings*, 3-65. New York: Cambridge University Press.
12. Dumais, S.T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1), pp. 188–230.
13. Graesser, A. C. et al. (1992). Mechanisms that generate questions. In: Lauer, T. et al. (eds.), *Questions and Information Systems*. Erlbaum, Hillsdale.
14. 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, Carnegie Mellon University.
15. Hoffart, J., Suchanek, F., Berberich, K. & Weikum, G. (2013). YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia, *Special issue of the Artificial Intelligence Journal.*
16. Huang, A. (2008). Similarity measures for text document clustering. In: *Proceedings of the 6th New Zealand Computer Science Research Student Conference*, pp. 49–56.
17. Jouault, C., & Seta, K. (2014). Content-dependent question generation for History learning in semantic open learning space. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, pp. 300-305.
18. Kunichika, H., Katayama, T., Hirashima, T. & Takeuchi, A. (2001): Automated Question Generation Methods for Intelligent English Learning Systems and its Evaluation. In: *Proceedings of the International Conference on Computers in Education*, pp. 1117-1124.
19. Le, N. T., Nguyen, N. P, Seta, K. & Pinkwart, N. (2014). Automatic Question Generation for Supporting Argumentation. *Vietnam Journal of Computer Science*, 1(2), pp. 117-127, Springer Verlag.
20. Le, N. T. & Pinkwart, N. (2015). Evaluation of a Question Generation Approach Using Open Linked Data for Supporting Argumentation. *Special Issue on Modeling, Management and Generation of Problems/Questions in Technology-Enhanced Learning, Journal of Research and Practice in Technology Enhanced Learning* (RPTEL).
21. Lin, L., Atkinson, R. K., Savenye, W. C., & Nelson, B. C. (2014). Effects of visual cues and self-explanation prompts: empirical evidence in a multimedia environment. *Interactive Learning Environments Journal.*
22. 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), pp. 101-124.
23. Mahdisoltani, F., Biega, J. & Suchanek, F. M. (2015). YAGO3: A Knowledge Base from Multilingual Wikipedias. In: *Proceedings of the Conference on Innovative Data Systems Research* (CIDR 2015).
24. Miller, G.A. (1995): WordNet: A lexical database for English. *Communications of the ACM*, 38(11), pp. 39-41.
25. Morgan, N. & Saxton, J. (2006). *Asking better questions*. Makhma, ON: Pembroke Publishers.
26. Mostow, J. & Chen, W. (2009). Generating Instruction Automatically for the Reading Strategy of Self-questioning. In: *Proceeding of the Conference on AI in Education*, pp.465-472
27. Mostow, J. & Beck, J. E. (2007). When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. In: B. Schneider & S.-K. McDonald (eds.), *Scale-Up in Education*, Rowman & Littlefield Publishers, pp. 183 - 200.
28. Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Journal Artificial Intelligence*, vol. 193, pp. 217-250.

29. Otero, J. & Graesser, A. C. (2001). PREG: elements of a model of Question Asking. *Journal Cognition and Instruction*, 19(2), pp. 143-175.
30. Pate, R. T. & Bremer, N. H. (1967). Guiding learning through skillful questioning. *Elementary School Journal*, 67, pp. 417-422.
31. Rigotti, E., & Greco Morasso, S. (2010). Comparing the Argumentum Model of Topics to Other Contemporary Approaches to Argument Schemes: The Procedural and Material Components. *Argumentation* 24(4), pp. 489-512.
32. Rothstein, D. & Santana, L. (2014). Teaching students to ask their own questions. *Havard Education Letter*, 27(5).
33. Schreiber, J. E. (1967). Teacher's question-asking techniques in social studies. *Doctoral dissertation, University of Iowa*, No. 67-9099.
34. Suchanek, F. M., Kasneci, G. & Weikum, G. (2007). YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In: *Proceedings of the International World Wide Web Conference*, pp. 697-706, ACM.
35. Tenenberg, J. & Murphy, L. (2005). Knowing What I Know: An Investigation of Undergraduate Knowledge and Self-Knowledge of Data Structures. *Computer Science Education*, 15(4), pp. 297-315.
36. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transaction on Information Systems*, 26(3), 13:1–13:37.
37. Yu, F.Y. & Pan, K.J (2014). The Effects of Student Question-Generation with Online Prompts on Learning. *Educational Technology & Society*, 17(3).