LearningQ: A Large-scale Dataset for Educational Question Generation

Guanliang Chen[1], **Jie Yang**[2], **Claudia Hauff**[1], **Geert-Jan Houben**[1]
[1]Delft University of Technology, [2]University of Fribourg
{guanliang.chen, c.hauff, g.j.p.m.houben}@tudelft.nl, jie@exascale.info

Abstract

We present LearningQ, a challenging educational question generation dataset containing over 230K document-question pairs. It includes 7K instructor-designed questions assessing knowledge concepts being taught and 223K learner-generated questions seeking in-depth understanding of the taught concepts. We show that, compared to existing datasets that can be used to generate educational questions, LearningQ (i) covers a wide range of educational topics and (ii) contains long and cognitively demanding documents for which question generation requires reasoning over the relationships between sentences and paragraphs. As a result, a significant percentage of LearningQ questions ($\sim$30%) require higher-order cognitive skills to solve (such as *applying*, *analyzing*), in contrast to existing question-generation datasets that are designed mostly for the lowest cognitive skill level (i.e. *remembering*). To understand the effectiveness of existing question generation methods in producing educational questions, we evaluate both rule-based and deep neural network based methods on LearningQ. Extensive experiments show that state-of-the-art methods which perform well on existing datasets cannot generate useful educational questions. This implies that LearningQ is a challenging test bed for the generation of high-quality educational questions and worth further investigation. We open-source the dataset and our codes at `https://dataverse.mpi-sws.org/dataverse/icwsm18`.

Introduction

In educational settings, questions are recognized as one of the most important tools not only for assessment but also for learning (Prince 2004). Questions allow learners to apply their knowledge, to test their understanding of concepts and ultimately, to reflect on their state of knowledge. This in turn enables learners to better direct their learning effort and improve their learning outcomes. Previous research has shown that the number of questions learners receive about a knowledge concept is positively correlated with the effectiveness of knowledge retention (Bahrick et al. 1993). It is thus desirable to have large-scale question banks for every taught concept in order to better support learners.

Designing a suitably large set of high-quality questions is a time-consuming and cognitively demanding task. Instructors need to create questions of varying types (e.g., open-ended, multiple choice, fill-in-the-blank), varying cognitive skill levels (e.g., applying, creating) and varying knowledge dimensions (e.g., factual, conceptual, procedural) that are preferably syntactically different yet semantically similar in order to enable repeated testing of a knowledge concept. To ease instructors' burden, *automatic question generation* has been proposed and investigated by both computer scientists and learning scientists to automate the question creation process through computational techniques (Mitkov, An Ha, and Karamanis 2006; Rus and Arthur 2009; Rus and Lester 2009; Heilman and Smith 2010).

Typically, automatic question generation has been tackled in a rule-based manner, where experienced teachers and course instructors are recruited to carefully define a set of rules to transform declarative sentences into interrogative questions (Wang, Hao, and Liu 2007; Adamson et al. 2013; Heilman and Smith 2010). The success of these rule-based methods is heavily dependent on the quality and quantity of the handcrafted rules, which rely on instructors' linguistic knowledge, domain knowledge and the amount of time they invest. This inevitably hinders these methods' ability to scale up to a large and diverse question bank.

Data-driven methods, deep neural network based methods in particular, have recently emerged as a promising approach for various natural language processing tasks, such as machine translation, named entity recognition and sentiment classification. Inspired by the success of these works, Du, Shao, and Cardie treated question generation processes as a sequence-to-sequence learning problem, which directly maps a piece of text (usually a sentence) to a question. In contrast to rule-based methods, these methods can capture complex question generation patterns from data without handcrafted rules, thus being much more scalable than rule-based methods. As with most data-driven approaches, the success of neural network based methods is largely dependent on the *size* of the dataset as well as its *quality* (Rajpurkar et al. 2016).

Exiting datasets, such as SQuAD (Rajpurkar et al. 2016) and RACE (Lai et al. 2017), though containing a large number of questions (e.g., 97K questions in SQuAD), are not suitable for question generation in the learning context. Instead of being aimed at educational question generation, these datasets were originally collected for reading comprehension tasks. They are often limited in their coverage of

This is the preprint version of the paper accepted at AAAI 2018.

Table 1: Examples of document-question pairs.

Source	***Doc*ument-*Q*uestion pairs**
SQuAD	*Doc:* ... after Heine's German birthplace of Düsseldorf had rejected, allegedly for anti-Semitic motives ... *Q:* Where was Heine born?
RACE	*Doc:* ... There is a big supermarket near Mrs. Green's home. She usually ... *Q:*: Where is the supermarket?
LearningQ	*Doc:* ... gases have energy that is proportional to the temperature. The higher the temperature, the higher the energy the gases have. The crazy thing is that at the same temperature, all gases have the same energy ... *Q:* If you were given oxygen (molecular mass = 18 AMU) and hydrogen (molecular mass = 1 AMU) at the same temperature and pressure, which has more energy?

topics—the questions in SQuAD for example, were generated by crowdworkers based on a limited number (536) of Wikipedia articles. More importantly, these questions seek factual details and the answer to each question can be found as a piece of text in the source passages; they do not require higher-level cognitive skills to answer them, as exemplified by the SQuAD and RACE example questions in Table 1. We speculate, as a consequence, question generators built on these datasets cannot generate questions of varying cognitive skill levels and knowledge dimensions that require a substantial amount of cognitive efforts to answer, which unavoidably limits the applicability of the trained question generators for educational purpose.

Our Contributions. To address these problems, we present LearningQ, which consists of more than 230K document-question pairs collected from mainstream online learning platforms. LearningQ does not only contain questions designed by instructors (7K) but also questions generated by students (223K) during their learning processes, e.g., watching videos and reading recommended materials. It covers a diverse set of educational topics ranging from computing, science, business, humanities, to math. Through both quantitative and qualitative analyses, we show that, compared to existing datasets, LearningQ contains more diverse and complex source documents; moreover, solving the questions requires higher-order cognitive skills (e.g., *applying*, *analyzing*). Specifically, we show that most questions in LearningQ are relevant to multiple source sentences in the corresponding document, suggesting that effective question generation requires reasoning over the relationships between document sentences, as shown by the LearningQ question example in Table 1. Finally, we evaluate both rule-based and state-of-the-art deep neural network based question generation methods on LearningQ. Our results show that methods which perform well on existing datasets cannot generate high-quality educational questions, suggesting that LearningQ is a challenging dataset worth of significant further study.

To the best of our knowledge, LearningQ is the first large-scale dataset for educational question generation. It provides a valuable data source for studying cross-domain question generation patterns. The distinct features of LearningQ make it a challenging dataset for driving the advances of automatic question generation methods, which will benefit learning and possibly also other domains where automatic generation of questions are necessary, e.g., conversational agents (Vinyals and Le 2015) that are expected to ask meaningful question so as to engage users.

Related Work

Question Generation

Automatic question generation has been envisioned since the late 1960s (Ross 1967). It is generally believed by learning scientists that the generation of high-quality learning questions should be based on the foundations of linguistic knowledge and domain knowledge, and thus they typically approach the task in a rule-based manner (Wang, Hao, and Liu 2007; Adamson et al. 2013; Heilman and Smith 2010; Mitkov and Ha 2003). Such rules mainly employ syntactic transformations to turn declarative sentences into interrogative questions (Chomsky 1973). For instance, Mitkov and Ha (2003) generated multiple-choice questions from documents by employing rules of term extraction. Based on a set of manually-defined rules, Heilman and Smith (2010) produced questions in a overgenerate-and-rank manner where questions are ranked based on their linguistic features. These methods, however, are intrinsically limited in scalability: rules developed in certain subjects (e.g., introductory linguistics, English learning) cannot be easily adapted to other domains; the process of defining rules requires considerable efforts from experienced teachers or domain experts. More importantly, manually designed rules are often incomplete and do not cover all possible document-question transformation patterns, thus limiting rule-based generators to produce high-quality questions.

Entering the era of large-scale online learning, e.g., Massive Open Online Courses (Pappano 2012), the demand for automatic question generation has been increasing rapidly along with the largely increased number of learners and online courses accessible to them. To meet the need, more advanced computational techniques, e.g., deep neural network based methods, have been proposed by computer scientists (Du, Shao, and Cardie 2017; Du and Cardie 2017). In the pioneering work by Du, Shao, and Cardie (2017), an encoder-decoder sequence learning framework (Sutskever, Vinyals, and Le 2014) incorporated with the global attention mechanism (Luong, Pham, and Manning 2015) was used for question generation. The proposed model can automatically capture question-asking patterns from the data, without relying on any hand-crafted rules, thus has achieved superior performance to rule-based methods in terms of both scalabilty and the quality of the generated questions.

These methods, however, have only been tested on datasets that were originally collected for machine reading comprehension tasks. Noticeably, these datasets contain a very limited number of useful questions for learning, as we will show in the following sections. Therefore, it remains an open question how deep neural network methods perform in processing complex learning documents and generating desirable educational questions.

Datasets for Question Generation

Several large-scale datasets have been collected to fuel the development of machine reading comprehension models, including SQuAD (Rajpurkar et al. 2016), RACE (Lai et al. 2017), NewsQA (Trischler et al. 2016), TriviaQA (Joshi et al. 2017), NarrativeQA (Kočiskỳ et al. 2017), etc. Though containing questions, all of these datasets are not suitable for educational question generation due to either the limited number of topics (Rajpurkar et al. 2016; Lai et al. 2017) or the loose dependency between documents and questions, i.e., a document might not contain the content necessary to generate a question and further answer the question. More importantly, most questions in these datasets are not specifically designed for learning activities. For example, SQuAD questions were generated by online crowdworkers and are used to seek for factual details in source documents; TriviaQA questions were retrieved from online trivia websites. An exception is RACE, which was collected from English examinations designed for middle school and high school students in China. Though collected in a learning context, RACE questions are mainly used to assess students' knowledge level of English, instead of other skills or knowledge of diverse learning subjects.

Depending on different teaching activities and learning goals, educational questions are expected to vary in cognitive complexity, i.e., requiring different levels of cognitive efforts from learners to solve. Ideally, an educational question generator should be able to generate questions of all cognitive complexity levels, e.g., from low-order recalling factual details to high-order judging the value of a new idea. This requires the dataset for training educational question generators to contain questions of different cognitive levels. As will be presented in our analysis, LearningQ covers a wide spectrum of learning subjects as well as cognitive complexity levels and is therefore expected to drive forward the research on automatic question generation.

Data Collection

Data Sources

To gather large amounts of useful learning questions, we initially explored several mainstream online learning platforms and finally settled on two after having considered the data accessibility and the quantity of the available questions as well as the corresponding source documents. Concretely, we gathered LearningQ data from the following two platforms:

TED-Ed[1] is an education initiative supported by TED which aims to spread the ideas and knowledge of teachers and students around the world. In TED-Ed, teachers can create their own interactive lessons, which usually involve lecture videos along with a set of carefully crafted questions to assess students' knowledge. Lesson topics range from humanity subjects like arts, language and philosophy to science subjects like business, economics and computer technology. Typically, a lesson, covering a single topic, includes one lecture video, and lasts from 5 to 15 minutes. Due to the subscription-free availability, TED-Ed has grown into one of the most popular educational communities and serves millions of teachers and students every week. As questions in TED-Ed are created by instructors, we consider them to be high-quality representations of testing cognitive skills at various levels (e.g., the LearningQ question in Table 1 is from TED-Ed). We use TED-Ed as the major data source to collect instructor-crafted learning questions.

Khan Academy[2] is another popular online learning platform. Similar to TED-Ed, Khan Academy also offers lessons to students around the world. Compared to TED-Ed, the lessons are targeted at a wider audience. For example, the math subjects in Khan Academy cover topics from kindergarten to high school. In addition, the lessons are organized in alignment with typical school curriculum (from the easier to the more advanced) instead of being an independent collection of videos as is the case in TED-Ed. Another distinction between the two platforms is that, Khan Academy allows learners to leave posts and ask questions about the learning materials (i.e., lecture videos and reading materials) during their learning. For instance, the chemistry course *Quantum numbers and orbitals*[3] includes one article (titled *The quantum mechanical model of the atom*) and three lecture videos (titled *Heisenberg uncertainty principle*, *Quantum numbers* and *Quantum numbers for the first four shells*) and learners can ask questions about any of them. More often than not, learners' questions express their confusion about the learning material—e.g., "How do you convert Celsius to Calvin?"—and thus are an expression of learners' knowledge gaps that need to be overcome in order to master the learning material. We argue that these questions can promote in-depth thinking and discussion among learners, thus complementing instructor-designed questions. We use those learner-generated questions as part of LearningQ.

We implemented site-specific crawlers for both Khan Academy and TED-Ed and collected all available questions and posts in English as well as their source documents at both platforms that were posted on or before December 31, 2017, resulting in a total of 1,146,299 questions and posts.

Question Classification for Khan Academy

Compared to instructor-designed questions collected from TED-Ed, learner-generated posts in Khan Academy can be of relatively low quality for our purposes since they are not guaranteed to contain a question (a learner may for example simply express her appreciation for the video) or the contained question can be off topic, lack the proper context, or be too generic. Examples of high- and low-quality questions are shown in Table 2.

[1] https://ed.ted.com/

[2] https://www.khanacademy.org/

[3] https://www.khanacademy.org/science/chemistry/electronic-structure-of-atoms/orbitals-and-electrons/

Originally, we gathered a total of 953,998 posts related to lecture videos and 192,301 posts related to articles from Khan Academy. To distill useful learning questions from the collected posts, we first extracted sentences ending with a question mark from all of the posts, which resulted in 407,723 such sentences from posts on lecture videos and 66,100 on reading material. To further discriminate useful questions for learning from non-useful ones, we randomly sampled 5,600 of these questions and recruited two education experts to annotate the questions: each expert labeled 3,100 questions (600 questions were labelled by both experts to determine the inter-annotator agreement) in a binary fashion: useful for learning or not. Based on the labelled data, we trained a convolutional neural network (Kim 2014) on top of pre-trained word embeddings (Mikolov et al. 2013) to classify the remaining Khan Academy questions. In the following, we describe the labelling process in more details.

Table 2: Examples of useful (marked with $\surd$) and non-useful questions from Khan Academy. S/H/M/C/E/T denote Science, Humanities, Math, Computing, Economics and Test Preparation, respectively.

ID	Questions	Topic	Label
a)	What is the direction of current in a circuit?	S	$\surd$
b)	Why can't voltage-gated channels be placed on the surface of Myelin?	S	$\surd$
c)	Is there a badge for finishing this course?	E	
d)	Have you looked on your badges page to see if it is one of the available badges?	T	
e)	Why do each of them have navels?	H	
f)	Does it represent phase difference between resistance and reactance?	S	
g)	What should the graph look like for higher voltages?	S	$\surd$
h)	What if some of the ideas come from different historical perspectives, giving inaccurate information?	H	
i)	What if the information is wrong ?	M	
j)	Can someone please help me?	C	
k)	Could you be more specific ?	T	
l)	Are you asking what geometric means?	M	
m)	Are you talking about the frequency?	E	
n)	What programming language or how much of coding I need to know to start learning algorithms here?	C	
o)	Can I do algorithms or should I do programming first?	C	

Question Annotation. We consider a user-generated question to be as useful for learning when all of the following conditions hold: (i) the question is *concept-relevant*, i.e., it seeks for information on knowledge concepts taught in lecture videos or articles; (ii) the question is *context-complete*, which means sufficient context information is provided to enable other learners to answer the question; and (iii) the question is not generic (e.g., a question asks for learning advice). To exemplify this, two concept-relevant learning questions are shown in Table 2 (*a* and *b*), accompanied by two concept-irrelevant ones (*c* and *d*). Question *e* and *f* in the same table are also concept-relevant. However, as they don't provide enough context information, e.g., lack of references for "they" and "it", we consider them as non-useful. As a counter-example, we consider question *g* in the table as useful since the reference for "the graph" can be easily inferred. This comes in contrast to question *h* and *i*, where the references for "the idea" and "the information" are too vague thus failed to to provide sufficient context information. Finally, generic questions expressing the need for help (*j* and *k*), asking for clarification (*l* and *m*) or general learning advice (as exemplified by *n* and *o*), are not useful for learning the specific knowledge concepts.

Annotation & Classification Results. Of the 5,600 annotated questions, we found 3,465 (61.9%) to be useful questions for learning. The inter-annotator agreement reached a Cohen's Kappa of 0.82, which suggests a substantially coherent perception of question usefulness by the two annotators. To understand the performance of the classifier trained on this labeled dataset, we randomly split the dataset into a training set of size 5,000, a validation set of size 300, and a test set of size 300. We iterated the training and evaluation process 20 times to obtain a reliable estimation of classification performance. Results show that the model reaches an accuracy of 80.5% on average (SD=1.8%), suggesting that the classifier can be confidently applied for useful/non-useful question classification. With this classifier, we retain about 223K unique useful questions in Khan Academy, which will be used for our following analysis.

Final Statistics of LearningQ

An overall description of LearningQ is shown in Table 3 (rows 1—4). As a means of comparison, we also provide the same statistics for the popular question generation datasets (though not necessarily useful for education and learning) SQuAD and RACE. Compared to these two datasets, LearningQ (i) consists of about 230K questions (versus 97K in SQuAD and 72K in RACE) on nearly 11K source documents; (ii) contains not only useful educational questions carefully designed by instructors but also those generated by learners for in-depth understanding and further discussion of the learning subject; (iii) covers a wide range of educational subjects from two mainstream online learning platforms. To highlight the characteristics of LearningQ, we also include SQuAD and RACE in the data analysis presented next.

Data Analysis

The complexity of questions with respect to the required cognitive skill levels and knowledge dimensions is a crucial property that can significantly influence the quality of questions for learning (LW et al. 2001). We thus believe that this factor should be studied when building efficient question generators. However, to our knowledge, there is no

Table 3: Descriptive features and statistics of LearningQ and the datasets in comparison.

Row	Feature Type	Features	SQuAD	RACE	TED-Ed	Khan Academy	
						Video	Article
1.	Basic statistics	Creator	Crowdworker	Instructor	Instructor	Learner	Learner
2.		# Unique documents	20,958	27,933	1,102	7,924	1,815
3.		# Unique questions	97,888	72,547	7,612	201,273	22,585
4.		# Avg. questions / document	4.67	2.60	6.91	25.40	12.44
5.	Document & question length	# Avg. words / document	134.84	322.88	847.64	1370.83	1306.55
6.		# Avg. sentence / document	4.96	17.63	42.89	73.51	49.85
7.		# Avg. words / sentence of documents	27.17	18.31	19.76	18.65	26.21
8.		# Avg. words / question	11.31	11.51	20.07	16.72	17.11
9.		# Avg. sentence / question	1.00	1.03	1.41	1.00	1.00
10.	Entities	# Avg. entities /document	10.24	9.75	17.66	14.55	47.38
11.		# Avg. entities /question	0.92	0.53	0.66	0.29	0.44
12.		% Entity words in question	8.10	4.58	3.29	1.72	2.54
13.	Readability	Document readability	45.82	73.49	64.08	76.54	55.15
14.		Question readability	67.23	51.00	66.32	72.15	69.04

work attempting to characterize this property of questions in datasets for question generation.

In this section, we characterize the cognitive complexity of questions in LearningQ and other existing question generation datasets along several dimensions: *(i)* low-level document and question attributes related to cognitive complexity (Wood 1986; Yang et al. 2016), e.g., the number of sentences or words per document or per question; *(ii)* document and question properties that can affect human perception of cognitive complexity, which include topical diversity, document and question readability (Collins-Thompson 2014; Sweller and Chandler 1994), etc.; and *(iii)* cognitive skill levels in accordance with Bloom's Revised Taxonomy (LW et al. 2001).

Document & Question Lengths

Table 3 (rows 5—9) presents statistics on document and question lengths. It can be observed that while, on average, the number of words per sentence in the documents of LearningQ are not larger than in SQuAD/RACE, documents from both TED-Ed and Khan Academy are more than twice as longer than those from SQuAD and RACE. In particular, SQuAD documents are on average nearly ten times shorter than Khan Academy documents. The same observation holds for the questions in LearningQ, where question length is twice as long as that of SQuAD and RACE. Compared with those in Khan Academy, documents in TED-Ed are shorter. This is mainly due to the fact that TED-Ed encourages shorter videos on a single topic .

Topics, Interrogative Words, and Readability

To gain an overview of the topics, we applied Named Entity Recognition to obtain statistics on the entities. The results are shown in rows 10 and 11 of Table 3. To gain more insights into the semantics of the documents and questions, we report the most frequent terms (after stopword removal) in Table 4 across both documents and questions. In order to gain insights into the type of questions, we separately consider interrogative terms (such as who or why) in the rightmost part of Table 4 by keeping most stopwords but filtering out prepositions and definite articles.

We observe in Table 3 that documents in LearningQ on average contains 160% more entities than SQuAD and RACE, which is expected because LearningQ documents are longer. Yet, the number of entities in LearningQ questions are not significantly larger than SQuAD and RACE. In particular, questions in SQuAD contain 40% more entities than those in LearningQ. This is despite the fact that SQuAD documents are shortest overall, as we showed earlier. To eliminate the influence of question lengths and refine the analysis, we further observe that the percentage of entities among all the words (row 12) in SQuAD questions is higher than that in LearningQ questions. The same observation holds when comparing RACE with LearningQ. These observations imply that, on the one hand, documents in LearningQ are more complex with respect to the number of involved entities; on the other hand, fewer questions related to entities, i.e., fewer factual questions, exist in LearningQ than the other datasets.

This interpretation is also supported by the top-k words shown in Table 4. We observe that while both documents and questions in SQuAD favor topics related to time and location (e.g., *time*, *year*, *century*, *city*, *state*), all data sources in LearningQ have fewer questions on these topics; more often in LearningQ questions, we find abstract words such as *mean*, *difference*, *function*, which are indicative of higher level cognitive skills being required. In line with this observation, we note that more interrogative words seeking factual details such as *who* and *when* rank high in the list of starting words of questions in SQuAD, while questions in LearningQ sources start much more frequently with *why*. This suggests that answering LearningQ questions often requires a deeper understanding of learning materials. Interestingly, one can observe in TED-Ed questions (in the middle part of the table) frequent words such as *think* and *explain*,

Table 4: Top words in documents and questions and top interrogative words of quiestions in LearningQ and the datasets in comparison. Words pertinent to a specific data source platforms are in bold. KA represents Khan Academy.

Top Words in Documents				**Top Words in Questions**				**Top Starting Words of Questions**			
SQuAD	RACE	TED-Ed	KA	SQuAD	RACE	TED-Ed	KA	SQuAD	RACE	TED-Ed	KA
new	people	like	**going**	year	**passage**	think	**number**	what	what	what	what
city	say	people	**let**	**type**	according	following	know	who	which	which	how
time	time	time	right	use	following	people	**Sal**	how	according	how	why
world	like	know	time	new	**author**	**explain**	**mean**	when	why	why	**is**
use	**day**	make	**equal**	**city**	**writer**	use	use	which	how	if	if
state	new	**way**	say	people	people	time	like	where	when	when	**can**
century	**school**	call	**plus**	call	know	like	right	why	if	who	**do**
united	make	world	**negative**	time	text	make	difference	according	who	according	when
war	year	think	**minus**	**war**	**learn**	different	**negative**	whose	where	**explain**	**are**
know	world	different	think	**located**	**probably**	world	**equation**	if	list	where	**would**

which explicitly ask learners to process learning materials in a specific way. These required actions are directly related to learning objectives defined by Bloom's Revised Taxonomy, as we will analyze later. In addition to the above, another interesting observation from Table 4 is that learners frequently ask questions for the clarification of videos using words such as *know*, *Sal* (the name of the instructor who initially created most videos at Khan Academy in the early stage of the platform), and *mean*.

Readability. Readability is an important document and question property related to learning performance. Table 3 (rows 13—14) reports the Flesch readability scores of documents and questions in the compared datasets (Collins-Thompson 2014). A piece of text with larger a Flesch readability score indicates it is easier to understand. Questions found alongside both Khan Academy videos and article possess similar readability scores, despite of the different sources. This confirms our previous finding on the similarity between the two subsets of Khan Academy data. We therefore do not distinguish these two subsets in the following analyses focused on questions.

Cognitive Skill Levels

It is generally accepted in educational research that a good performance on assessment questions usually translates into "good learning" (Hibbard and others 1996). We first use Bloom's Revised Taxonomy to categorize the questions according to the required cognitive efforts behind them (LW et al. 2001). The taxonomy provides guidelines for educators to write learning objectives, design the curriculum and create assessment items aligned with the learning objectives. It consists of six learning objectives (requiring different cognitive skill levels from lower order to higher order):

- **Remembering**: questions that are designed for retrieving relevant knowledge from long-term memory.
- **Understanding**: questions that require constructing meaning from instructional messages, including oral, written and graphic communication.
- **Applying**: questions that ask for carrying out or using a procedure in a given situation.
- **Analyzing**: questions that require learners to break material into constituent parts and determine how parts relate to one another and to an overall structure or purpose.
- **Evaluating**: questions that ask for make judgments based on criteria and standards.
- **Creating**: questions that require learners to put elements together to form a coherent whole or to re-organize into a new pattern or structure.

To exemplify, we select one question example for each category that we collected from TED-Ed and Khan Academy, as shown in Table 5. Among the different learning objectives defined by Bloom's Revised Taxonomy, *analyzing* is an objective closely related to the task of automatic question generation. *analyzing* questions require the learner to understand the relationships between different parts of the learning material. Existing question generation methods (Du, Shao, and Cardie 2017), however, can usually only take one sentence as input. To cope with *analyzing* questions, state-of-the-art methods first need to determine the most relevant sentence in the learning material, which is then used as input to the question generator. This inevitably limits the ability of trained question generators to deliver meaningful *analyzing* questions covering multiple knowledge concepts scattered in the source documents. To understand the complexity of the LearningQ questions specifically from the point of view of training question generators, we also include in our analyses an exploration of the proportion of questions at various Bloom levels that require knowledge from multiple source document sentences.

Data Annotation. To facilitate our analysis, we recruited two experienced instructors to label 200 randomly sampled questions from each of the compared datasets according to Bloom's Revised Taxonomy. The Cohen's Kappa agreement score between the two annotators reached 0.73, which is a substantial agreement. In a second labeling step, we labelled the selected questions with their sentence(s) based on which they are generated.

Comparative Results. Table 6 shows the results of question classification according to Bloom's Revised Taxonomy. SQuAD only contains *remembering* questions, suggesting that it is the least complex dataset among all compared datasets in terms of required cognitive skill levels.

Table 5: Question Examples of Different Bloom' Revised Taxonomy Level in TED-Ed and Khan Academy.

Taxonomy	TED-Ed Examples	Khan Academy Examples
Remembering	How big is an atom?	What is a negative and a positive feedback in homeostasis?
Understanding	Why do some plankton migrate vertically?	Why can't voltage-gated channels be placed on the surface of myelin?
Applying	What kind of invention would you make with shape memory materials if you could get it in any form you wanted?	If i double the area and take the half of the fraction, do I get the same result?
Analyzing	Why are cities like London, Tokyo, and New York facing shortages in burial ground space?	Why did sea levels drop during the ice age?
Evaluating	Mansa Musa is one of many African monarchs throughout the continent's rich history. Yet, the narratives of only a few kings and queens are featured in television and movies. Analyze and evaluate why you think that this is the case, then create two ideas for how we can work to bring more positive awareness of the history of Africa's ancient and contemporary kings and queens to students today.	Will all the cultures merge into one big culture, due to the fading genetic distinctions?
Creating		Can somebody please explain to me what marginal benefits is and give me some examples?

In general, we note a trend of decreasing percent of *remembering* questions (and increasing percentage of *understanding* questions) from SQuAD, RACE, to TED-Ed and Khan Academy. We can conclude that questions in LearningQ demand higher cognitive skills than those in SQuAD and RACE. Interestingly, among the two different LearningQ sources, we can observe that there are more *understanding* and *applying* questions in Khan Academy than in in TED-Ed, while there are more *evaluating* and *creating* questions in TED-Ed than in Khan Academy. This shows the inherent differences related to the corresponding learning activities between instructor-designed questions and learner-generated questions. The former is mainly used for assessment purpose and thus contains more questions of higher-order cognitive skill levels; the latter is generated during students' learning process (e.g., watching lecture videos and reading recommended materials) and is usually used to seek for a better understanding of the learning material. Note that 26.42% Khan Academy questions questions were labelled as either irrelevant or unknown due to being not useful for learning or missing enough context information for the labeller to assign a Bloom category. This aligns with the accuracy of the useful question classifier we reported in the data collection section.

In Table 7 we report the results of our source sentence(s) labeling efforts. From the statistics of # words in source sentences, we can observe an increasing requirement for reasoning over multiple sentences from SQuAD and RACE to TED-Ed and Khan Academy. Compared to the 98.5% of single sentence related questions in SQuAD, questions in TED-Ed (Khan Academy) are related to 3.53 (6.65) sentences on average in source documents. In particular, Table 7 (the last row) further shows that a large portion of the questions in LearningQ, especially in Khan Academy, cannot be answered by simply relying on the source document, as exemplied by the *evaluating/creating* question from TED-Ed in Table 5 and thus require external knowledge to generate.

Table 6: Distribution of Bloom's Revised Taxonomy Labels.

	SQuAD	RACE	TED-Ed	Khan Academy
Remembering	100	82.19	61.86	18.24
Understanding	0	18.26	38.66	55.97
Applying	0	0.46	9.79	12.58
Analyzing	0	8.22	14.95	15.09
Evaluating	0	1.37	4.12	1.89
Creating	0	0	1.55	0.63
Unknown/ Irrelevant	0	3.20	0	26.42

Experiments and Results

In this section, we conduct experiments to evaluate the performance of rule-based and deep neural network based methods in question generation using LearningQ. We aim to answer the following questions: 1) how effective are these methods at generating high-quality educational questions; 2) to what extent is their performance influenced by the learning topics; and 3) to what extent does the source sentence(s) length affect the question generation performance.

Experimental Setup

Comparison methods. We investigate a representative rule-based baseline and two state-of-the-art deep neural networks in question generation:

- **H&S** is a rule-based system which can be used to generate questions from source text for educational assessment

Table 7: Results of Source Sentence Labelling. # Words/Sent. denote the average words/sentences in the labelled source sentences. % ONE/MULTIPLE/EXTERNAL refer to the percentage of questions related to ONE single sentence, MULTIPLE sentences or require EXTERNAL knowledge to generate, respectively. KA denotes Khan Academy.

	SQuAD	**RACE**	**TED-Ed**	**KA**
# Words	32.39	46.02	76.57	128.23
# Sent.	1.01	2.87	3.53	6.65
% ONE	98.53	37.10	28.63	9.43
% MULTIPLE	1.47	62.90	52.42	23.27
% EXTERNAL	0	0	18.95	38.99

and practice (Heilman and Smith 2010). The system produces questions in a overgenerate-and-rank manner. We only evaluate the top-ranked question.

- **Seq2Seq** is a representative encoder-decoder sequence learning framework proposed for machine translation (Sutskever, Vinyals, and Le 2014). It automatically learns the patterns of transforming an input sentence to an output sentence based on training data.
- **Attention Seq2Seq** is the state-of-the-art method proposed in (Du, Shao, and Cardie 2017), which incorporates the global attention mechanism (Luong, Pham, and Manning 2015) into the encoder-decoder sequence learning framework during the decoding process. The attention mechanism allows the model to mimic humans problem-solving process by focusing on relevant information in the source text and using this information to generate a question.

Data Preparation. We use the NLTK tool (Bird and Loper 2004) to pre-process the dataset: lower-casing, tokenization and sentence splitting. To account for the fact that existing methods can only process a small number of sentences as input, for each question, we use the following strategy inspired by approaches for text similarity (Gomaa and Fahmy 2013) to locate the source sentences in the corresponding document most relevant to the question. If the target question contains a timestamp—e.g., "in 10:32, what does the Sal mean ..."—indicating the source sentence(s) location from which the target question is generated, we then choose that sentence as the starting sentence and compute the cosine similarity with the target question. We then go forwards and backwards in turns to determine whether including a nearby sentence would increase the cosine similarity between the target question and the source sentences. If yes, the nearby sentence is added. Otherwise, the search process stops. If a target question does not contain timestamp information, we select the sentence with largest cosine similarity to the question to start our search the same way as described above to locate the source sentences. Due to the vanishing gradient problem in recurrent neural networks (Hochreiter et al. 2001), we only keep data with source sentences containing no more than 100 words.

Notice that deep neural network based methods usually require a substantial amount training data. The quantity of instructor-crafted questions in TED-Ed is not sufficient (7K). We therefore train the selected methods only on learner-generated questions. Concretely, we first merge all of the questions posted by Khan Academy learners on both lecture videos and reading materials, then randomly select 80% for training, 10% for validation and 10% for testing. At the same time, we also use all of the instructor-crafted questions as a second test set to investigate how effective the models built on learner-generated questions are in delivering instructor-crafted questions.

Parameter Settings. We implemente the two neural network based methods on top of the OpenNMT system (Klein et al. 2017). In accordance with the original work (Sutskever, Vinyals, and Le 2014; Du, Shao, and Cardie 2017), Bi-LSTM is used for the encoder and LSTM for the decoder. We tune all hyper-parameters using the held-out validation set and select the parameters that achieve the lowest perplexity on the validation set. The number of LSTM layers is set to 2 and its number of hidden units is set to 600. The dimension of input word embedding is set to 300 and we use the pre-trained embeddings *glove.840B.300d* for initialization (Pennington, Socher, and Manning 2014). Model optimization is performed by applying Adam (Kingma and Ba 2014); we set the learning rate to 0.001 and the dropout rate to 0.3. The gradient is clipped if it exceeds 5. We train the models for 15 epochs in mini-batches of 64. When generating a question, beam search with a beam size of 3 is used and the generation stops when every beam in the stack produces the `<EOS>` (end-of-sentence) token.

Evaluation Metrics. Similar to (Du, Shao, and Cardie 2017), we adopt Bleu 1, Bleu 2, Bleu 3, Bleu 4, Meteor and $Rouge_L$ for evaluation. Bleu-n scores rely on the maximum n-grams for counting the co-occurrences between a generated question and a set of reference questions; the average of Bleu is employed as final score (Papineni et al. 2002). Meteor computes the similarity between the generated question and the reference questions by taking synonyms, stemming and paraphrases into account (Denkowski and Lavie 2014). $Rouge_L$ reports the recall rate of the generated question with respect to the reference questions based on the longest common sub-sequence (Lin 2004).

Results and Analysis

Results. Table 8 reports the performance of the selected methods on learner-generated questions from Khan Academy and instructor-designed questions from TED-Ed. We can observe that across all different evaluation metrics, the rule-based method H&S is outperformed by both deep neural network based methods. This confirms previous findings in the new context of learning that data-driven methods are a better approach for question generation. Among the two deep neural network based methods, Attention Seq2Seq consistently outperform Seq2Seq (p-value $< .001$, Paired t-test). This verifies that the attention mechanism is an effective approach for boosting the performance of educational question generation.

Table 8: Performance of rule-based and deep neural network based methods on LearningQ.

	Methods	Bleu 1	Bleu 2	Bleu 3	Bleu 4	Meteor	Rouge$_L$
Khan Academy	H&S	0.28	0.17	0.13	0.10	3.24	6.61
	Seq2Seq	19.84	7.68	4.02	2.29	6.44	23.11
	Attention Seq2Seq	24.70	11.68	6.36	3.63	8.73	27.36
TED-Ed	H&S	0.38	0.22	0.17	0.15	3.00	6.52
	Seq2Seq	12.96	3.95	1.82	0.73	4.34	16.09
	Attention Seq2Seq	15.83	5.63	2.63	1.15	5.32	17.69

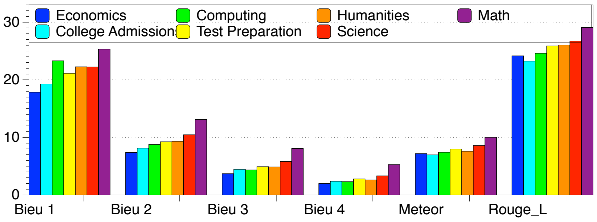
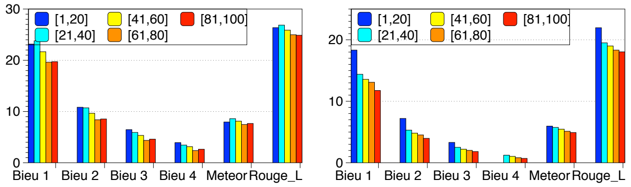


Figure 1: Results of question generation on different learning subjects in Khan Academy.



(a) Khan Academy

(b) TED-Ed

Figure 2: Results of question generation with different source sentence lengths..

By comparing the performance of the selected methods on Khan Academy and on TED-Ed, we find that the performance of rule-based method H&S varies across different evaluation metrics. The performance measured by Bleu scores are higher on learner-generated questions than on instructor-designed questions, while it is low as measured by Meteor and *Rouge*$_L$. On the other hand, deep neural network based methods consistently reach a higher performance on learner-generated questions than on instructor-designed questions. Considering the fact that recurrent networks are less effective in handling long sentences, this could be due to two reasons: 1) the majority of questions in TED-Ed are related to multiple sentences as we found (Table 7); and 2) the questions generated by learners are generally shorter than those designed by instructors (Table 3). In later analysis, we further describe how the length of source sentences would affect question generation performance.

The performance of the state-of-the-art methods is much lower on LearningQ than on existing datatsets. Attention Seq2Seq achieves a Bleu 4 score $>$ 12 and a Meteor score $>$ 16 on SQuAD, while on LearningQ it only achieves Bleu 4 scores of $<$ 4/$<$ 2 and Meteor scores of $<$ 9/$<$ 6 on learner-generated questions/instructor-designed questions, respectively. Similar results also hold for the other metrics.

Impacts of Subjects and Source Sentence Lengths. We now investigate the performance of Attention Seq2Seq in generating educational questions as affected by different subjects and different lengths of input source sentences.

The impact of the source document topic on question generation performance is shown for Khan Academy in Figure 1. We observe that question generation performance varies across subjects. In particular, Bleu 4 varies from <2 to >5 for learner-generated questions and from 0.38 to 0.92 for instructor-designed questions. Compared to Economics and College Admissions, question generation for Math and Science can usually achieve higher performance. Similar variation is also observed on TED-Ed. These results indicate that topics can affect question generation performance. Fully understanding the co-influence of topics and other document properties (e.g., difficulty) however requires more studies, which we leave to future work.

As we showed before (Table 7), educational questions are related to multiple source sentences in the documents. However, existing neural network methods usually take only one or two source sentences as input to generate questions. To further investigate the effectiveness of existing methods when taking source sentences of different lengths as input, we divide the testing set according to the length of their source sentences. The results are shown in Figure 2. In general, question generation performance decreases when the length of source sentences increases across all metrics for both Khan Academy and TED-Ed. This strongly suggests that the performance of the state-of-the-art method is significantly limited by long source sentences.

Conclusion and Future Work

We presented LearningQ, a large-scale dataset for educational question generation. It consists of 230K document-question pairs produced by both instructors and learners. To our knowledge, LearningQ is the first dataset that covers a wide range of educational topics and the questions require a full spectrum of cognitive skills to solve. Extensive evaluation of state-of-the-art question generation methods on LearningQ showed that LearningQ is a challenging dataset that deserves significant future investigation.

For future research, deep neural network based methods can be further enhanced by considering the relationships among multiple source sentences for question generation.

Acknowledgement

This research project is supported by i) the *Extension School* of the Delft University of Technology and ii) the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 683253/GraphInt).

References

Adamson, D.; Bhartiya, D.; Gujral, B.; Kedia, R.; Singh, A.; and Rosé, C. P. 2013. Automatically generating discussion questions. In *AIED*.

Bahrick, H. P.; Bahrick, L. E.; Bahrick, A. S.; and Bahrick, P. E. 1993. Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science* 4(5):316–321.

Bird, S., and Loper, E. 2004. NLTK: the natural language toolkit. In *ACL*.

Chomsky, N. 1973. Conditions on transformations.

Collins-Thompson, K. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* 165(2):97–135.

Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *SMT*.

Du, X., and Cardie, C. 2017. Identifying where to focus in reading comprehension for neural question generation. In *EMNLP*.

Du, X.; Shao, J.; and Cardie, C. 2017. Learning to ask: Neural question generation for reading comprehension. In *ACL*.

Gomaa, W. H., and Fahmy, A. A. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68(13).

Heilman, M., and Smith, N. A. 2010. Good question! statistical ranking for question generation. In *HLT-NAACL*.

Hibbard, K. M., et al. 1996. *Performance-Based Learning and Assessment. A Teacher's Guide.*

Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J.; et al. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; and Rush, A. M. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Kočiskỳ, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2017. The narrativeqa reading comprehension challenge. *TACL*.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

LW, A.; DR, K.; PW, A.; KA, C.; Mayer, R.; PR, P.; D. Raths, J.; and MC, W. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

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

Mitkov, R., and Ha, L. A. 2003. Computer-aided generation of multiple-choice tests. In *HLT-NAACL*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Pappano, L. 2012. The year of the mooc. *The New York Times* 2(12):2012.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Prince, M. 2004. Does active learning work? a review of the research. *Journal of engineering education* 93(3):223–231.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Ross, J. R. 1967. Constraints on variables in syntax.

Rus, V., and Arthur, C. G. 2009. The question generation shared task and evaluation challenge. In *The University of Memphis. National Science Foundation*.

Rus, V., and Lester, J. 2009. The 2nd workshop on question generation. In *AIED*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Sweller, J., and Chandler, P. 1994. Why some material is difficult to learn. *Cognition and instruction* 12(3):185–233.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2016. Newsqa: A machine comprehension dataset. *CoRR* abs/1611.09830.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wang, W.; Hao, T.; and Liu, W. 2007. Automatic question generation for learning evaluation in medicine. In *ICWL*.

Wood, R. E. 1986. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* 37(1):60 – 82.

Yang, J.; Redi, J.; Demartini, G.; and Bozzon, A. 2016. Modeling task complexity in crowdsourcing. In *HCOMP*.