Automatic True/False Question Generation for Educational Purpose

Bowei Zou[1]*, **Pengfei Li**[2]*, **Liangming Pan**[3], **Ai Ti Aw**[1]
[1]Institute for Infocomm Research, A*STAR, Singapore
[2]Nanyang Technological University, Singapore
[3]National University of Singapore, Singapore
{zou_bowei,aaiti}@i2r.a-star.edu.sg
pengfei.li@ntu.edu.sg
liangmingpan@u.nus.edu

Abstract

In field of teaching, true/false questioning is an important educational method for assessing students' general understanding of learning materials. Manually creating such questions requires extensive human effort and expert knowledge. Question Generation (QG) technique offers the possibility to automatically generate a large number of questions. However, there is limited work on automatic true/false question generation due to the lack of training data and difficulty finding question-worthy content. In this paper, we propose an unsupervised True/False Question Generation approach (TF-QG) that automatically generates true/false questions from a given passage for reading comprehension test. TF-QG consists of a template-based framework that aims to test the specific knowledge in the passage by leveraging various NLP techniques, and a generative framework to generate more flexible and complicated questions by using a novel masking-and-infilling strategy. Human evaluation shows that our approach can generate high-quality and valuable true/false questions. In addition, simulated testing on the generated questions challenges the state-of-the-art inference models from NLI, QA, and fact verification tasks.

1 Introduction

For educational purposes, questioning not only assesses the acquisition of knowledge, but also reinforces the engagement and critical thinking of learners during effective teaching, which in turn enables learners to clearly guide their learning efforts and enhance their skills (Prince, 2004). With the ever-growing educational content on the internet and the increasing popularity of online tutoring applications during the COVID-19 pandemic, an automatic question creation process becomes a key technique to reduce the efforts in manually constructing questions and facilitate adaptive learning.

Text-based question generation for education aims to produce legible and pedagogically-salient questions from a given textual content to provide meaningful learning experiences, where the answer to the question can be found or derived from the content. Earlier QG models generate simple questions based on manually constructed rules (Rus et al., 2012; Lindberg et al., 2013; Lee, 2016). However, such questions often lack linguistic diversity and contain much ungrammatical or nonsensical content (Kurdi et al., 2020). Recently, with the development of deep learning and question answering (QA) techniques, the studies of QG have shifted towards neural question generation (NQG) which utilizes deep neural networks to generate more fluent and diverse questions (Pan et al., 2019). Depending on the QA datasets used for training, various types of questions can be generated such as span-based questions (Du et al., 2017; Gao et al., 2019), multiple-choice questions (Chung et al., 2020), and multi-hop questions (Pan et al., 2020; Su et al., 2020). However, due to the limitation of the current QA corpus, most of the generated questions focus on finding the information presented in the passage. Moreover, the majority of NQG models are used for improving QA or dialogue systems instead of for educational purposes (Duan et al., 2017; Sachan and Xing, 2018; Pan et al., 2021).

Among various types of educational-purposed questions, true/false (T/F) questions can yield valid assessments directly, simply, and efficiently (Ebel, 1970), which is useful to evaluate if the learners hold any misconceptions about the given material. In this paper, we take the approach of defining the T/F question as a declarative sentence (statement)[1], rather than an interrogative sentence like that in BoolQ (Clark et al., 2019). So far, automatically generating such type of questions is relatively less explored. Lee (2016) developed a system where the original sentences in passage are

* Equal contribution

[1]See more examples in Section 3.4.

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 61 - 70
July 15, 2022 ©2022 Association for Computational Linguistics

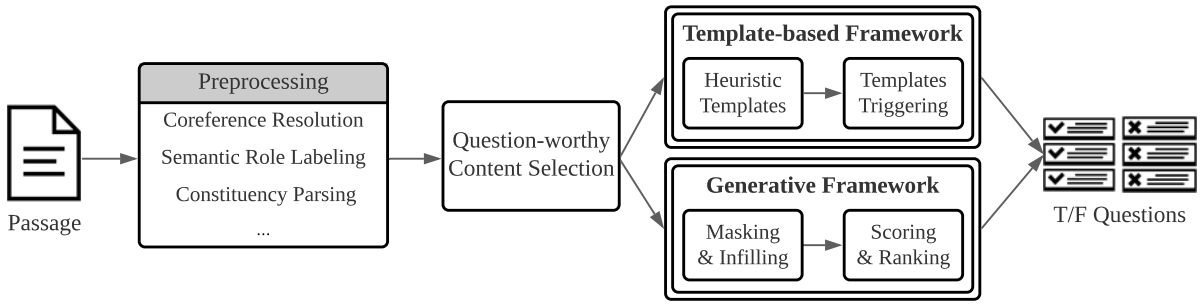

Figure 1: Overall architecture of TF-QG.

used as true questions and the false questions are generated by replacing the keywords with their antonyms or adding a negative keyword. Killawala et al. (2018)'s method was also based on simple syntactic templates. However, the quality of the generated questions is not good enough for assessment due to 1) the lack of training data and difficulty of finding good testing points from a given passage, and 2) the high occurrence of grammatical and semantic errors (Zhang and Bansal, 2019).

In this paper, we propose an unsupervised True/False Question Generation approach (TF-QG) for assessing the reading comprehension ability of English learners. TF-QG leverages both a traditional template-based method and a recently developed generative language model to generate high-quality T/F questions from a given passage. In the *template-based framework*, various NLP techniques are used for creating heuristic templates to test certain knowledge such as lexical, syntactic, and coreference understanding. In the *generative framework*, we propose a novel masking-and-infilling strategy to generate more flexible and complicated questions such as inferential questions that require deeper understandings of the passage. Specifically, to yield questions with valid testing points, we design several templates and mask selection protocols to select question-worthy contents from the passage. Then, the pretrained language model with text infilling objective is used to generate new statements based on both the prior knowledge and the context of the passage. Finally, we design a novel scoring mechanism to score and rank the generated questions based on their conciseness and relevance to the passage.

Extensive human evaluation shows that TF-QG is able to generate high-quality T/F questions containing both factoid and inferential content. In addition, simulated experiments on the generated T/F questions challenge the state-of-the-art NLI, QA, and fact verification systems, which indicates that these questions are difficult to some extent.

To summarize, our main contributions are:

- We propose an unsupervised system for T/F question generation with the educational purpose of testing students' reading comprehension ability. The question-worthy contents are selected by our designed templates and mask selection protocols targeting various testing points. Such templates and protocols can be customized by educators based on test points, making it easier to incorporate into TF-QG without modifying or retraining the model.
- We propose a masking-and-infilling question generation strategy that enables the system to generate more linguistically diverse and semantically complicated T/F questions.
- TF-QG provides a domain-independent solution for constructing a large-scale T/F reading comprehension dataset. Both human evaluation and simulated tests on reasoning tasks show the reasonableness and difficulty of the generated T/F questions.

2 TF-QG Model

Given a passage as reading material, TF-QG aims to generate T/F questions to test learners' understanding of the passage. The overall architecture is shown in Figure 1. The passage is first preprocessed to obtain the basic syntactic and semantic information (Section 2.1). Then, two unsupervised frameworks including the template-based framework (Section 2.2) and the generative framework (Section 2.3) are applied to generate T/F questions targeting the question-worthy contents in the passage. The question-worthy contents are selected according to our designed templates/protocols in the two frameworks which will be described in the respective sections.

2.1 Passage Pre-processing

We first conduct coreference resolution to resolve pronouns to their corresponding antecedents and gather antecedents representing the same concept into a *coreference set*. Then we implement semantic role labeling (SRL) and put the semantic roles of the same subject (Arg0) into respective *SRL sets*. The constituency parsing tree for each sentence is obtained by a syntactic parser. Finally, we extract *numeral sets* from the passage, each set contains instances of "number + quantifier" (e.g., "200 meters") with the same quantifier. Our implementations are based on the AllenNLP library (Gardner et al., 2017).[2]

2.2 Template-based Framework

To assess learners, intuitively, the generated T/F questions should be sufficiently similar to some fragments about the passage, but different from the passage at a pedagogically meaningful point. Although there be various definitions of what one might consider valuable test points, this paper focuses on the areas that we thought were most likely to be relevant to language learning and understanding. To this end, we design the following heuristic templates to generate T/F questions by selecting and modifying the question-worthy content in the given passage.

- **Coreference substitution template (Coref)** If a pronoun is more than one sentence away from its antecedent, we replace the pronoun with its antecedent to generate a true question. Besides, the pronoun is replaced with an irrelevant antecedent in the coreference set to generate a false question.
- **Coordination modification template (Coord)** From the constituency parsing tree, we find noun coordination structures in the form of "NP_1 CC NP_2" or "NP_1, NP_2, ..., CC NP_k".[3] Then we randomly select a NP_i ($i \in 1, ..., k$) node and use the templates "... only NP_i ..." and "... no NP_i ..." to generate false questions.
- **SRL modification template (SRL)** If there are same semantic role types in an SRL set, we exchange the two semantic roles into each other's sentences to generate two false questions.
- **Synonym/Antonym substitution template (Synonym/Antonym)** When we find an adjective or an adverb in a short sentence (<15 words), the word is replaced with its synonym or antonym from WordNet[4] to generate a true question or a false question, respectively.
- **Negation modification template (Negation)** If a sentence contains a verbal negation or a word from the negative cue list extracted from Bioscope (Vincze et al., 2008), we remove the negative word and take the rest of the sentence as a false question.
- **Number modification template (Num)** If there is more than one element in a numeral set, we randomly exchange two of them into each other's original sentences, to generate two false questions.
- **Definition modification template (Def)** If an appositive clause fits the pattern "... NP_1 <comma> NP_2 ...", we generate a corresponding true question as "NP_1 <copula> NP_2.".
- **Simplification rule** To make the question more concise and focus on the key information, we remove 1) the constituency structures "SBAR" and "IN+S", 2) the contents between two commas (parenthesis), and 3) the constituency structures "PP" and "ADVP" at the beginning of the question.

Each of the above heuristic templates is activated independently and repeatedly if its conditions are met. These templates aim to test the learners' understanding of the passage from different aspects: Coref, Num, and Def templates focus on the understanding of context meaning, number, and definition, respectively; Synonym/Antonym templates test learners' lexical understanding while Coord, SRL, and Negation template tests syntactic or semantic understanding.

Note that the above templates are customizable, i.e. educators could easily add new heuristic templates to TF-QG for specific teaching or testing purposes with. In addition, an advantage of the template-based framework is that it can generate T/F questions while determining whether their answers are true or false. On the other hand, the limitation of this template-based framework is that it requires educators to 1) know which types of language capabilities of the learners they would like to test and specify the test points (this is related to the educational process and difficult to be replaced by models), and 2) know the formulation of the fundamental NLP tasks, to smoothly convert the language test points to the templates with extra effort only once.

[2] https://allennlp.org

[3] NP: noun phrase; CC: coordinating conjunction.

[4] https://wordnet.princeton.edu

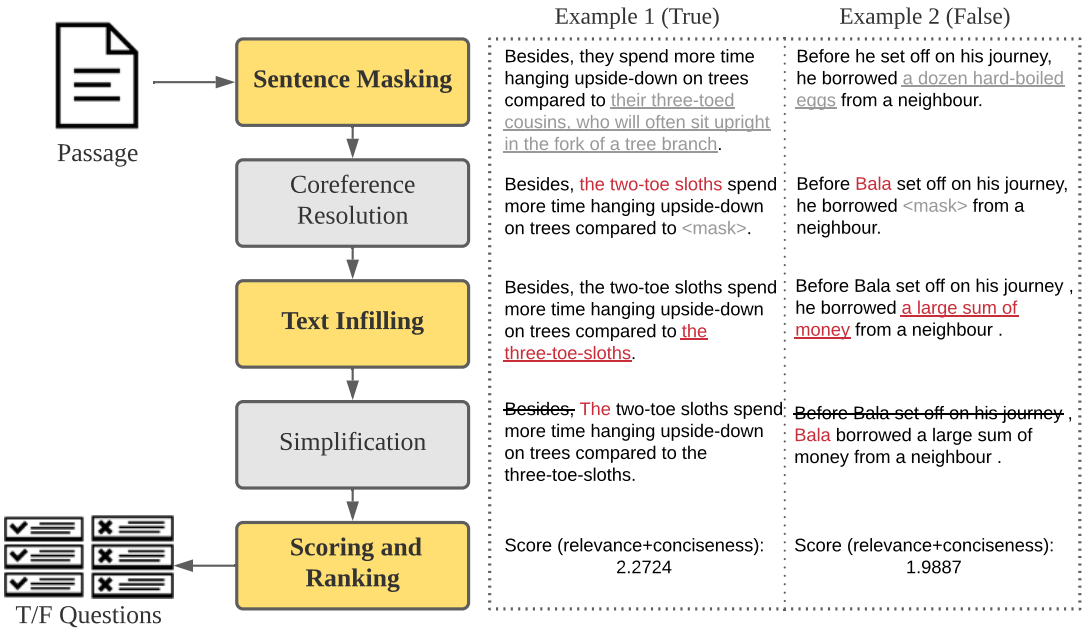

Figure 2: Generative framework of TF-QG. The processes and data-flow are shown on the left; two examples with step-by-step transformations are shown on the right.

2.3 Generative Framework

Generative framework aims to generate more flexible and complicated T/F questions. As shown on the left of Figure 2, the highlighted masking-and-infilling and scoring-and-ranking are the main components of our model. Two examples with step-by-step transformations are shown on the right. Example 1 is a true question generated from an expository passage, whereas Example 2 is a false question generated from a narrative passage. In the following, we describe each process in detail.

2.3.1 Sentence Masking

To pick question-worthy content from the passage and facilitate the generation of T/F questions, we design the following mask selection protocols.

- **Semantic role masking**. Mask the arguments of a predicate in the sentence based on the SRL results.
- **Subordinate clause masking**. Mask the part in a subordinate clause that follows a subordinating conjunction such as "that", "when", "since", etc.
- **Prepositional phrase masking**. Mask the part in a prepositional phrase that follows a preposition. We only consider the phrase with more than two words.
- **Adversative clause masking**. Mask the adversative clause in the sentence. The adversative relation is identified by the keywords such as "although", "but", etc. We also convert the keywords into coordinating conjunctions ("so"/"and") in order to generate false statements.
- **Declarative clause masking**. Mask the simple declarative clause after a preposition or subordinating conjunction (i.e. "IN+S").
- **Number masking**. Mask numbers. "one" is excluded since it is often used for other purposes.

These protocols identify the key information in the passage. Such information is replaced with a special <mask> token that represents a missing span in the sentence. More examples of the T/F questions generated from the above-mentioned mask selection protocols are provided in the case study in Section 3.4.

Coreference Resolution To improve clarity, the first-appeared pronouns in the sentence are replaced with their corresponding antecedents.

2.3.2 Text Infilling

To generate T/F questions from the masked sentences, we perform a text infilling task aiming to predict the missing span of text which are consistent with the preceding and subsequent text. We utilize a pretrained language model BART (Lewis et al., 2020) to perform text infilling, which is a Transformer-based denoising autoencoder pretrained on large text corpus with text infilling as a training objective. Hence, it has good capabilities of reconstructing a corrupted text by fitting the most suitable text to the missing span.

Criteria	Rating	Score	Description
Fluency (grammatical correctness)	bad	1	Not readable due to grammatical errors.
	fair	2	Contain few grammatical errors but not affect the readability too much.
	good	3	Free from grammatical errors.
Semantic (clarity and logical correctness)	bad	1	Have obvious logical/common-sense problem or indecipherable.
	fair	2	Have some semantic ambiguities.
	good	3	Semantically clear.
Relevance (to the passage)	bad	1	Totally irrelevant.
	fair	2	Part of the question is irrelevant.
	good	3	Relevant.
Answerability	bad	1	Not answerable.
	fair	2	Not sure about the correct answer.
	good	3	Can be answered by the right answer.
Difficulty	factoid	1	Can be inferred from a single sentence in the passage.
	inferential	2	Requires deeper understanding of the passage or longer context.

Table 1: Human evaluation metrics with description.

To make the generated text more relevant to the passage, we provide two sentences before and after the masked sentence as context to BART model. The model predicts the missing span based on both the context of the passage and the prior knowledge learned during language modeling. We also perform beam search with beam width 5 to obtain the top-5 outputs with the highest probabilities.

Simplification To make the question more concise, we perform the same simplification process as in the template-based framework by removing the auxiliary components of the sentence.

2.3.3 Scoring and Ranking

We propose a scoring mechanism to automatically evaluate and rank the generated questions based on their conciseness and relevance.

$$S = \frac{1}{1 + e^{-0.3(l_t - l_g)}} + \frac{R_l + R_c + R_s}{|g|}$$

The first term is the conciseness score where l_t and l_g are the lengths of the original and generated sentence, respectively. The second term is the relevance score where R_l is lexical relevance score measuring the number of overlapping words between the generated texts and the passage; R_c and R_s are conceptual and semantic relevance scores measuring the number of generated words that are conceptually and semantically relevant to the masked words. We use ConceptNet (Speer et al., 2017) to obtain the concept-relevant terms of the masked words, and FrameNet (Ruppenhofer et al., 2006) to obtain the semantic frames of both generated words and masked words. $|g|$ is a normalization term that counts the number of generated words.

Finally, we choose the question with the highest score from the beam search results. Then we rank all the questions generated from the passage and select the top-scoring questions as the final T/F questions.

3 Experimentation

3.1 Settings and Evaluation Metrics

Since there is no standard dataset available for automatic evaluation, we conduct human evaluation on the generated T/F questions. We randomly select 20 well-edited English passages from the quiz materials at a level of elementary education as our test set, which contains both expository writings (e.g., descriptive articles) and narrative writings (e.g., stories and diaries) on topics of general interest. For each passage, we collect all questions generated by the template-based framework and up to 20 questions generated by the generative framework. Finally, from the selected 20 passages, we obtain 401 questions in total, an average of 20 questions per passage.

Due to the educational nature of our purpose, we recruit three annotators with educational backgrounds to rate the produced questions. The annotators were first asked to read the passage, and then give judgments for fluency, semantic, relevance, answerability, and difficulty, as shown in Table 1. From the ratings given by the three annotators, we take the majority vote as the final ratings. In case of a tie, we choose the average rating (i.e. "fair"). In addition, for the results from the generative frame-

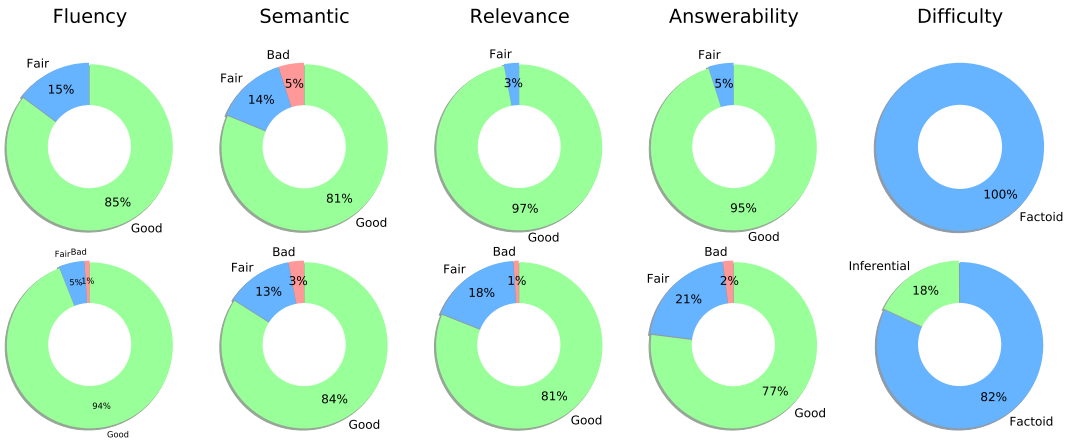

Figure 3: Human evaluation results of the T/F questions generated by TF-QG. Top row and bottom row show the results from template-based framework and generative framework, respectively.

Criteria	Template-based	Generative
Fluency	0.772	0.870
Semantic	0.723	0.710
Relevance	0.861	0.630
Answerability	0.812	0.620
Difficulty	0.881	0.813
Answer	/	0.725

Table 2: Annotator agreement. Scores denotes Randolph's kappa (Randolph, 2005) that measures the agreement from multiple annotators.

work, we also ask the annotators to label the answer (T/F) of the questions. Table 2 shows the inter-rater agreement, which indicates that all the annotations have substantial ($0.6 < \kappa \leq 0.8$) or almost perfect ($\kappa > 0.8$) agreement.

3.2 Experimental Results

The human evaluation results are presented in Figure 3. It is observed that the majority (>80%) of the questions generated by TF-QG have good fluency, semantic, relevance, and answerability. Hence, the questions are promising to be directly used for the educational purpose of assessing language learners' reading comprehension ability. However, we also observe that the generated questions have lower scores on the difficulty rating. All of the questions generated by the template-based framework are factoid, and only 18% of the questions from the generative framework are inferential. Finding such answers does not require too complicated reasoning efforts. Hence, we argue that the current method is still a long way from generating more complex questions, and this paper has played a role in exploring this direction.

For the template-based framework, templates offer the ability to produce questions lightly coupled with the exact wording of the original text. The results show that our TF-QG model can generate much more relevant questions with good answerability than the generative framework (relevance rating) since all generated questions are closely related to the passage, which makes the templates easy to leverage human linguistic expertise to produce questions tailored to specific educational content. In addition, the template-based framework also has the advantage that the answers are given explicitly since templates are designed for different types (true/false) of answers. However, the rigid transformations by templates may cause more grammatical (fluency rating) and logical (semantic rating) problems.

For the generative framework, the fluency and semantic of the questions are improved due to the benefits of language modeling. The two properties are crucial since if the generated questions do not satisfy such requirements, learners may easily be misled and frustrated, which reduces questions' pedagogical value. Besides, the syntactic and content of the questions are more flexible, enabling our model to generate more complicated questions. The human evaluation shows that our generative framework is able to produce inferential questions (18%) to test student's comprehensive understanding of the passage. However, due to the flexibility of generated content, the question may be irrelevant to the passage and hence their answerability may be affected.

Tasks	Dev	full	1sent	3sent	5sent
NLI	86.1	55.9	**66.2**	61.6	59.0
BoolQ	80.4	48.5	57.2	55.4	53.9
BoolQ_d	77.0	47.7	54.6	53.1	51.0
FEVER	95.3	50.3	52.8	51.0	52.6

Table 3: True/false reading comprehension accuracy (%). BoolQ_d: the questions are converted to declarative sentences. Dev: the performance on the development set of the fine-tuning tasks.

3.3 True/False Reading Comprehension

To further evaluate the difficulty of the questions generated by our model, we create a simulated task of true/false reading comprehension, which aims to test the capability of NLP models to answer T/F questions. To this end, we first construct a test set (TFQA) using the questions generated from the generative framework of TF-QG. Then, we ask the annotators to label the answers (True/False) of the questions. After removing the questions with bad answerability, the TFQA test set contains 210 false questions and 178 true questions. Finally, we test the performance of the state-of-the-art natural language inference (NLI), QA, and fact verification models on TFQA in a zero-shot transfer learning way. Specifically, we fine-tune a pretrained BERT (Devlin et al., 2019) model on various related tasks/datasets, including the NLI task with SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), the bool QA task with BoolQ (Clark et al., 2019), and the fact verification task with FEVER (Thorne et al., 2018).

For BoolQ, we use two strategies to make the task similar to ours: 1) convert our questions to interrogative sentences during inference; 2) convert BoolQ questions to declarative sentences during fine-tuning. Besides using the full passage ("full") as input, we also test the performance using the question-related sentence ("1sent") and the sentence with contexts including one sentence ("3sent") and two sentences ("5sent") before and after the sentence.

Experimental results are shown in Table 3. Although the model can achieve near state-of-the-art performances on the fine-tuning tasks ("Dev"), the best accuracy on the TFQA test set is only 66.2%. This demonstrates that the questions generated by our model are challenging. To obtain better performance, more sophisticated models and training data are required under supervised settings. Although the point is outside the scope of this paper, our approach does offer the NLP community a possibility to construct a T/F question answering dataset.

3.4 Case Study

We present a case study on a passage about "Yellowstone National Park". The questions generated by our TF-QG model are shown in Table 4. We show only one question for each template/protocol due to the space limitation.

Generally, the questions generated from the templates meet our goal of testing certain knowledge such as coreference, lexical, and definition understanding. However, since the template does not refer to the contextual information when substituting synonyms, Question 1 is not fluent due to the wrong wording. Question 3 shows the advantages of the template-based framework on the testing target that aims to distinguish concepts. In the original passage, Old Faithful is described as a "geyser", while in the question, it is stated as another approximate concept "hot spring". Question 4 also fulfills the test goal of concept understanding, which distinguishes concepts between "Celsius" and "Fahrenheit", although the generated question merely swaps the numbers. Regarding other test points, Question 2 provides a simple verbal negative case. Question 5 tests both pronoun understanding and vocabulary comprehension.

The questions generated by the generative framework are more flexible and challenging. Many questions require inferring from longer context and they are useful to test learners' comprehensive understanding of the passage, such as Question 6-9. In particular, the generative model supplements Question 9 with the information that "boiling water comes from geyser", which can only be obtained from the above description. Such questions can well examine the learner's understanding of contextual consistency and cohesion. However, some questions are hard to answer due to bad coreference resolution or irrelevant content generated as shown in Question 10.

In general, we observe that the generated T/F questions can be effectively targeted to test many teaching inspection points. Currently, although these generated questions are relatively simple, they are sufficient for usage in some scenarios, such as reading comprehension tests for primary school

Yellowstone National Park is in the United States of America. It became the first National Park in 1872. [1]There are geysers and hot springs at Yellowstone. There are also many animals like elk, bison, sheep, grizzly bears, black bears, moose, coyotes, and more at Yellowstone. More than 3 million people visit Yellowstone each year. [2]During the winter, visitors can ski, go snowmobiling or join tours there. [6]Visitors can see steam and water from the geysers. During other seasons, visitors can go horse-riding, boating, fishing or take nature trails and tours. [3,7]Most visitors want to see Old Faithful, a very predictable geyser at Yellowstone. Visitors can check a schedule to see the precise time that Old Faithful is going to erupt. There are many other geysers and bubbling springs in the area. [8]Great Fountain Geyser erupts every 11 hours up to a height of 67 metres. Excelsior Geyser produces 4,000 gallons of boiling water each minute! [4,9]Boiling water is 100 degrees Celsius, or 212 degrees Fahrenheit – that's very hot! People also like to see the Grand Prismatic Spring. It is the largest hot spring in the park. [5]It has many beautiful colors, which are caused by bacteria in the water. [10]These are forms of life that have only one cell. Different bacteria live in different water temperatures. Visiting Yellowstone National Park can be a week–long vacation or more. It is beautiful, and there are activities for everyone.

No.	Framework	Template/Protocol	True/False Question
1	Template	Synonym	There are geysers and spicy springs at Yellowstone. (F)
2	Template	Coord+Negation	During the winter, visitors cannot ski. (F)
3	Template	Def+Coord	Old Faithful is a very predictable hot spring at Yellowstone. (F)
4	Template	Num	Boiling water is 212 degrees Celsius. (F)
5	Template	Coref+Antonym	The Grand Prismatic Spring has many ugly colors, which are caused by bacteria in the water. (F)
6	Generative	Preposition	Visitors can see steam and water from Yellowstone's geysers and hot springs. (F)
7	Generative	Semantic Role Arg1	Most visitors want to see Old Faithful when it is erupting. (T)
8	Generative	Semantic Role Arg0	Yellowstone National Park is home to the world's largest geyser,Yellowstone Geyser,which erupts every 11 hours up to a height of 67 metres.(F)
9	Generative	Number	The temperature of the geyser water is about 100 degrees Celsius, or 212 degrees Fahrenheit - that's very hot! (T)
10	Generative	Subordinate	These are forms of life that live on the surface of water. (?)

Table 4: Questions generated from a passage describing "Yellowstone National Park". The text where each question is generated from is highlighted in the passage with the corresponding number. The masked text in the generative framework is indicated using underline. "(?)" means unanswerable.

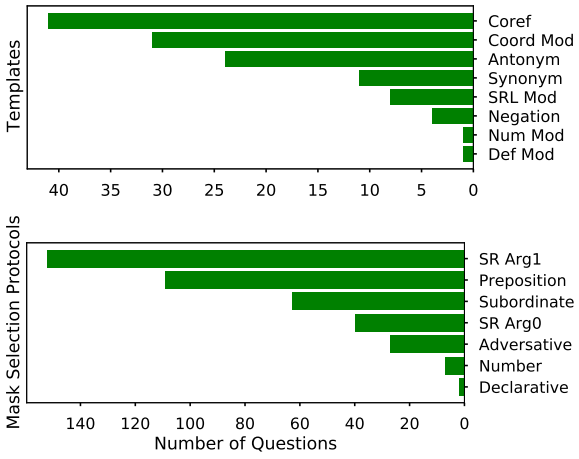

Figure 4: Number of questions generated from different templates (top) and mask selection protocols (bottom). "SR Arg0" and "SR Arg1" denote the semantic role masking protocol with the subject and object of the predicate being masked, respectively.

students or second language learners, or language education-oriented speech dialogue test systems.

3.5 Statistics of Templates and Protocols

We also study the frequency of different templates and mask selection protocols triggered by our TF-QG model. Figure 4 shows the number of questions generated from different templates/mask selection protocols based on the 20 testing passages. We can see that coreference, coordination, and antonym are the most frequently triggered templates for the template-based framework. For the generative framework, semantic role masking and prepositional phrase masking are the most frequently triggered mask selection protocols. The different numbers of the template- or protocol-triggered samples describe the distribution of the corresponding test points in the selected passages. Although we carefully selected different types of passages (including expository articles, stories, and diaries), more passages from different domains and genres still need to be explored to further verify the robustness of our proposed model on T/F question generation.

Besides, it is observed that the generative framework can the generate more questions than template-based framework in total. In fact, the masking-and-infilling approach allows the generative framework to produce an infinite number of questions, but the question quality still has to be considered. We currently pick questions by the

generative confidence of the model. In future work, a more pedagogical question selection approach should be taken into account, such as which protocols should be selected in terms of practical quiz objectives, and which protocols are more suitable for generating inferential or challenging questions for different genres.

4 Conclusion

In this paper, we propose an automatic true/false question generation approach, which provides a feasible scheme for large-scale generation of educational content. Two unsupervised frameworks including template-based framework and generative framework are proposed to select question-worthy contents from the passage and generate high-quality questions. The novel masking-and-infilling strategy enables our model to generate more flexible and complicated true/false questions.

In future work, we will focus on how to design templates and mask selection protocols to match with pedagogically valuable test points proposed by domain experts. In addition, we will perform controlled lab or online studies to measure students' learning gains after studying the content generated by TF-QG. Furthermore, we expect to deploy the proposed approach on real educational platforms, including an interactive language learning and assessment system (for students), and a question generation assistance system (for teachers), to measure how much the approach could reduce the workload of educators in practical application scenarios.

Acknowledgements

The research has been supported by Institute of Infocomm Research of A*STAR (CR-2021-001). We thank the anonymous reviewers for their valuable and constructive feedback.

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642. Association for Computational Linguistics (ACL).

Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A bert-based distractor generation scheme with multi-tasking and negative answer training strategies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4390–4400.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874.

Robert L Ebel. 1970. The case for true-false test items. *The School Review*, 78(3):373–389.

Yifan Gao, Lidong Bing, Wang Chen, Michael R Lyu, and Irwin King. 2019. Difficulty controllable generation of reading comprehension questions. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4968–4974. AAAI Press.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Akhil Killawala, Igor Khokhlov, and Leon Reznik. 2018. Computational intelligence framework for automatic quiz question generation. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Jae-Young Lee. 2016. Dynamic relocation of true-false questions using ready-made arrays with random numbers. *International Journal of Software Engineering and Its Applications*, 10(8):91–100.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 105–114.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Unsupervised multi-hop question answering by question generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5866–5880, Online. Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475.

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

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater κfree): An alternative to fleiss' fixed-marginal multirater kappa. In *Presented at the Joensuu Learning and Instruction Symposium*, volume 2005.

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, and Jan Scheffczyk. 2006. Framenet ii: Extended theory and practice. berkeley. *CA: International Computer Science Institute*.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2012. A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse*, 3(2):177–204.

Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 629–640.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020. Multi-hop question generation with graph convolutional network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4636–4647.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1–9.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2495–2509.