Textual Entailment based Question Generation

Takaaki Matsumoto[1,2], **Kimihiro Hasegawa**[3], **Yukari Yamakawa**[1], and **Teruko Mitamura**[1]

[1]Carnegie Mellon University
[2]SOC Corporation
[3]Kobe University
{tmatsumo, yukariy, teruko}@andrew.cmu.edu, ljnjzbo417@gmail.com

Abstract

This paper proposes a novel question generation (QG) approach based on textual entailment. Many previous QG studies transform a single sentence into a question directly. They need hand-crafted templates or generate simple questions similar to the source texts. As a novel approach to QG, this research employs two-step QG: 1) generating new texts entailed by source documents, and 2) transforming the entailed sentences into questions. This process can generate questions that need the understanding of textual entailment to solve. Our system collected 1,367 English Wikipedia sentences as QG source, retrieved 647 entailed sentences from the web, and transformed them into questions. The evaluation result showed that our system successfully generated non-trivial questions based on textual entailment with 53% accuracy.

1 Introduction

Question generation (QG) is a practical application field of natural language generation. One important objective of QG in education is cultivating students' reading comprehension skills.

Many studies have been done on QG by transforming a single sentence into a question. Heilman and Smith (2010) researched QG based on syntactic parsing which is characterized by overgenerating and scoring. Mazidi and Tarau (2016) generated questions based on dependency parsing. Woo et al. (2016) studied QG based on dependency and semantic role labeling.

Their systems can generate relatively simple but grammatical questions. Suppose the following sentence is picked up from the website[1] .

1. *Kawabata won the Nobel Prize in Literature for his novel "Snow Country".*

Using the sentence above as a source, Heilman's system[2] generated the following question.

2. *Did Kawabata win the Nobel Prize in Literature for his novel "Snow Country"?*

Although this question is grammatical, its educational effectiveness could be minimized, since students might not exert their reading comprehension skills due to the similarity between the generated question and the original sentence. Questions generated by these QG methods are often quite similar to the original sentences.

Some researchers have tried inference QG with templates. Labutov et al. (2015) studied a QG system that utilizes ontology and templates developed by crowd workers. Chinkina and Meurers (2017) built a conceptual QG system using hand-crafted pattern matching templates. Although the templates in these studies may need more work, the generated questions are more complicated than those by transforming the single sentence.

Our research proposes a novel QG approach based on textual entailment. In contrast to the existing studies that directly generate questions from sources, our system firstly generates new sentences entailed by source texts and then transforms the entailed sentences into questions as shown in Figures 1 and 2. For example, we generate the following sentence entailed by the sentence (1).

3. *Kawabata is the writer of "Snow Country".*

Now we create a question for sentence (1) by transforming sentence (3) as follows.

[1] https://sites.google.com/site/ntcir11riteval/

[2] http://www.cs.cmu.edu/ ark/mheilman/questions/

Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG), pages 15–19,
Tilburg, The Netherlands, November 5 2018. ©2018 Association for Computational Linguistics

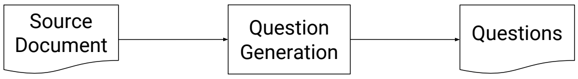

Figure 1: Existing QG Flow

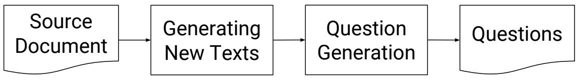

Figure 2: Proposed QG Flow

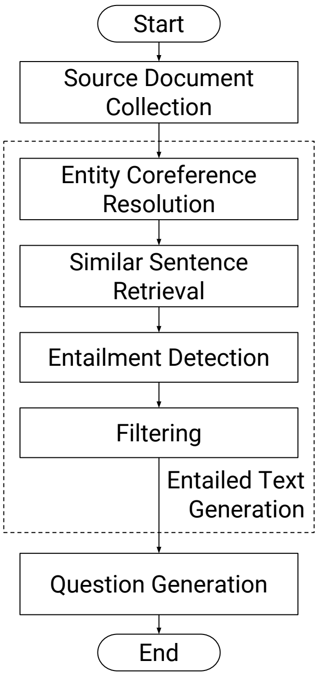
Start
Source Document Collection
Entity Coreference Resolution
Similar Sentence Retrieval
Entailment Detection
Filtering
Entailed Text Generation
Question Generation
End

Figure 3: System Flow of the Proposed Approach

4. *Is Kawabata the writer of "Snow Country"?*

This question requires students to utilize more reading comprehension skills than the question (2), because there is not a word "writer" in the sentence (1). Students need to infer that Kawabata is a writer from the phrase of "won the Nobel Prize in Literature" in the sentence (1) using world knowledge. This method enables us to generate questions that are not similar to the original sentences but need textual entailment inference to solve.

2 Proposed Method

Figure 3 illustrates the QG process of this research. We first collected source texts. Second, we retrieved new texts entailed by the source sentences. Finally, we generated questions based on the entailed sentences. In the subsections below, we describe the function of each module.

2.1 Source Document Collection

Source texts for QG were collected from English Wikipedia. To generate entailed sentences for each source sentence, sentence tokenization using spaCy[3] was applied to all the collected sentences.

One example sentence from "Taj Mahal" article in English Wikipedia was the following:

5. *It is regarded by many as the best example of Mughal architecture and a symbol of India's rich history.*

2.2 Entailed Text Generation

We generate entailed texts by applying entailment detection to similar texts retrieved from the web.

2.2.1 Entity Coreference Resolution

To search texts similar to the collected sentences effectively, the entity coreferences of the source texts were resolved by using neuralcoref[4]. Coreferent entities are often important keywords to search similar sentences.

For example, the entity coreference of the sentence (5) was resolved as follows:

6. *The Taj Mahal is regarded by many as the best example of Mughal architecture and a symbol of India's rich history.*

2.2.2 Similar Sentence Retrieval

We then retrieved similar sentences from the web for each sentence with entity coreference resolved (for example, retrieving sentence (3) from sentence (1) in Section 1). In order to select sentences similar to the original text, we employed spaCy's sentence embedding to measure the similarity of sentences.

The following sentences are examples of the retrieved sentences for sentence (6) in this step.

7. *India, the Taj Mahal is by common consent the finest example of Mughal Architecture.*

8. *The Taj Mahal is considered one of the finest specimen of the Mughal architecture.*

9. *The Taj Mahal incorporates and expands on design traditions of Persian and earlier Mughal architecture.*

2.2.3 Entailment Detection

To extract entailed sentences from the similar retrieved sentences, we applied the ESMI entailment

[3] https://spacy.io

[4] https://github.com/huggingface/neuralcoref

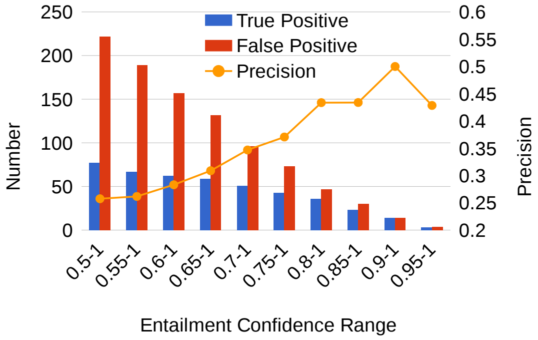

Figure 4: Preliminary Entailment Evaluation

detector (Chen et al., 2017), which was trained using MultiNLI (Williams et al., 2018). We employed the GloVe (Pennington et al., 2014) as the word embedding for the ESIM.

For sentences (7), (8), and (9), the entailment detector labeled "entailment (confidence 0.93)," "entailment (confidence 0.60)," and "neutral (confidence 0.86)," respectively. Sentences (7) and (8) were kept because they were labeled as entailment. However, we eliminated sentence (9) because of the neutral label.

2.2.4 Filtering

Filtering was applied to improve the answer existence accuracy of the generated questions. Our filtering metrics include entailment confidence, ROUGE-1, sentence similarity, and the word counts of the source sentences and the retrieved ones. Table 1 shows the thresholds we used.

For sentences (7) and (8), sentence (7) was kept because it met all the criteria. However, sentence (8) was excluded because it did not satisfy the entailment confidence criterion.

Entailment Confidence Although the entailment detector can classify similar sentences retrieved, we filtered some results of entailment detection to increase the precision. Figure 4 shows the preliminary results of human evaluation. We collected Wikipedia sentences and retrieved the sentences labeled as "entailment" by the ESIM. As can be seen, the precision increases in proportion to the minimum threshold of the entailment confidence. To improve the entailment detector precision, we used a high confidence as a threshold.

ROUGE-1 To control the ratio of word overlapping between the entailed sentences and the source sentences, ROUGE-1 (Lin, 2004) was used.

Sentence Similarity We used spaCy to calculate sentence similarity between the source sentences and the entailed ones, because ROUGE-1 cannot measure semantic similarity.

Table 1: Filtering Values

	Min.	Max.
Entailment confidence	0.9	1
ROUGE-1	0.2	0.7
Sentence similarity	0.5	1
Num. of retrieved sentence words	6	-
Num. of source sentence words	Num. of retrieved sentence words	-

The Word Counts of the Source Sentences and the Retrieved Ones We excluded too short sentences because they tend not to contain enough information.If a source sentence is too short, an entailed sentence would also be too short to make a question.

2.3 Question Generation

Questions based on textual entailment were generated by an existing QG tool. We chose Heilmen's QG system because it has been widely used as a baseline QG system in many papers (Woo et al. (2016) and Mazidi and Tarau (2016)). We picked up the top ranked yes/no question for each source sentence because Heilman's tool overgenerates questions.

For example, sentence (7) was transformed into the following question.

10. *Is the Taj Mahal by common consent the finest example of Mughal Architecture?*

3 Experiment

We collected 100 English Wikipedia abstracts as QG sources. We extracted 1,360 sentences by the sentence tokenization and their entity coreferences were resolved Then, we retrieved 61,330 similar sentences from the web (maximum 50 similar sentences per sentence). Maximum 30, 10 and 10 sentences were selected from the Google search results, English Wikipedia, and Simple Wikipedia, respectively. The entailment detector labeled 16,770 sentences as textual entailment with an argmax criterion, but 676 sentences remained after the filtering. We applied Heilman's QG system to them.

Answerable Examples:

1. Article: IQ

Source Sentence:

Unlike, for example, distance and mass, a concrete measure of intelligence cannot be achieved given the abstract nature of the concept of "intelligence".

H&S System's Yes/no Question:

Can distance and mass not be achieved given the abstract nature of the concept of ``intelligence" for example?

Our System's Question:

Is it problematic to claim that the intelligence quotient is a measure of intelligence?

Retrieved Sentence:

So, it is problematic to claim that the intelligence quotient is a measure of intelligence.

2. Article: Classical economics

Source Sentence:

These economists produced a theory of market economies as largely self-regulating systems, governed by natural laws of production and exchange (famously captured by Adam Smith's metaphor of the invisible hand).

H&S System's Yes/no Question:

Were largely self-regulating systems governed by natural laws of production and exchange?

Our System's Question:

Is the invisible hand a natural force that self regulates the market economy?

Retrieved Sentence:

The invisible hand is a natural force that self regulates the market economy.

3. Article: Taj Mahal

Source Sentence:

It is regarded by many as the best example of Mughal architecture and a symbol of India's rich history.

H&S System's Yes/no Question:

(No yes/no questions were generated)

Our System's Question:

Is the Taj Mahal by common consent the finest example of Mughal Architecture?

Retrieved Sentence:

India, the Taj Mahal is by common consent the finest example of Mughal Architecture.

Unanswerable Examples:

1. Article: Castle

Source Sentence:

Many castles were originally built from earth and timber, but had their defences replaced later by stone.

H&S System's Yes/no Question:

Were many castles originally built from earth and timber?

Our System's Question:

Were castles?

Retrieved Sentence:

Castles, whether made of mortared stone or earth and timber, were.

Error: Similar sentence retrieval failure

2. Article: Hydrogen

Source Sentence:

Hydrogen is a chemical element with symbol H and atomic number 1.

H&S System's Yes/no Question:

Is hydrogen a chemical element with symbol H and atomic number 1?

Our System's Question:

Is Fermium a chemical element?

Retrieved Sentence:

Fermium (symbol Fm) is a chemical element.

Error: Entailment detection failure

3. Article: Measles

Source Sentence:

Measles is an airborne disease which spreads easily through the coughs and sneezes of infected people.

H&S System's Yes/no Question:

Is Measles an airborne disease which spreads easily through the coughs and sneezes of infected people?

Our System's Question:

Does coughs, or sneezes spread through the air?

Retrieved Sentence:

When an infected person breathes, coughs, or sneezes, the virus spreads through the air.

Error: Entailment was OK but question generation failed.

Figure 5: Examples of the Questions from Our System

3.1 Discussion

We evaluated 150 out of 676 generated questions. The evaluation results suggested that our system successfully generated textually entailed questions with 53% accuracy. Figure 5 lists answerable and unanswerable examples of the generated questions. Tables 2, 3, and 4 show the grammaticality, textual entailment, and answer existence of the evaluated questions, respectively.

The positive examples shown in Figure 5 suggests that the proposed method successfully generated relatively complex questions compared with Heilman's tool. In the first positive example, for instance, students need to infer that IQ is "problematic" to measure intelligence by the phrase of "a concrete measure of intelligence cannot be achieved" in the source sentence.

The questions from our system relatively shared a few number of words with the source sentences compared to questions directly generated from the source sentences by Heilman's tool. We measured two mean scores of ROUGE-1 (1) between the source texts and our system's questions, and (2) between the source texts and the questions generated directly from the source sentences by Heilman's tool. The mean scores of ROUGE-1 were 0.76 and 0.36, respectively. This difference suggests that the questions from our system would require more reading comprehension skills than the questions from Heilman's tool.

Table 2: Grammaticality of Questions

	Number	Ratio
Ungrammatical	26	0.17
Grammatical w/ minor errors	33	0.22
Grammatical	91	0.61

Table 3: Entailment of Retrieved Sentences

	Number	Ratio
Not Entailed	66	0.44
Entailed	84	0.56

Table 4: Answer Existence of Questions

	Number	Ratio
Unanswerable	70	0.47
Answerable	80	0.53

As can be seen in Table 2, about 83% of the evaluated questions were grammatical or grammatical with minor errors. Out of 26 ungrammatical questions, 14 were due to the errors of Heilman's system and 12 due to the errors in the retrieval process.

The evaluations of textual entailment and answer existence (Tables 3 and 4) were similar to each other because most of the unanswerable questions were generated from not-entailed sentences. However, there are a few exceptions. The retrieved text of the third unanswerable example in Figure 5 was entailed by the source text, but Heilman's tool generated an unanswerable question.

4 Conclusion

In this paper, a new question generation method using textually entailed information is proposed. We implemented the question generation system that utilizes textual entailment and applied it to English Wikipedia abstracts. For 1,367 source sentences, our system generated 647 questions and more than half of the evaluated questions were answerable. In the future, we plan to develop a natural language generation method to generate entailed sentences based on given texts instead of retrieving entailed sentences from the web.

Acknowledgements

The authors would like to acknowledge Dr. Jun Araki for his advice on this research. This publication was partly made possible by grant NPRP-08-1337-1-243 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver. ACL.

Maria Chinkina and Detmar Meurers. 2017. Question generation for language learning: From ensuring texts are read to supporting learning. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 334–344.

Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617. Association for Computational Linguistics.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 889–898.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Karen Mazidi and Paul Tarau. 2016. Infusing nlu into automatic question generation. In *Proceedings of the 9th International Natural Language Generation conference*, pages 51–60.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1112–1122.

Simon Woo, Zuyao Li, and Jelena Mirkovic. 2016. Good automatic authentication question generation. In *Proceedings of the 9th International Natural Language Generation conference*, pages 203–206.