Combination of Statistical and Neural Approaches in the Japanese Question Generation System

Lasguido Nio Koji Murakami
Rakuten Institute of Technology, Rakuten Inc.
{lasguido.nio, koji.murakami}@rakuten.com

1 Abstract

Question generation is the task of generating a natural question from a given input sentence. This often requires hand-crafted templates or sophisticated NLP pipelines which require extensive labor and expertise to morphologically analyze the sentences and create the NLP framework. In order to simplify these labors, contrastive experiment between two types of sequence learning: statistical-based machine translation and attention-based sequence neural network have been done. Going further we investigate, compares, and observe the combination of both approaches for question generation task. Combination of both approaches is done based on the voting mechanism. This way we aim to overcome both approaches and establish a system that excels in terms of content quality and fluency according to a subjective human test.

2 Introduction

Language generation is attractive because its applications involve a vast amount of domains ranging from dialogue systems [14, 10], reading comprehension [15, 11, 5], to Frequently Asked Questions (FAQ) generation [18]. Question generation seems to be simple yet not many scientists explore it. Creating a good question is not a trivial task, and it is much harder to create a deep question rather than a factoid question. While factoid questions require an explicit memory response, deep questions require more profound thinking and recall [8]. In this study, we will limit ourselves to questions of factoid type that can be answered explicitly but are tedious to create.

Conventional approaches on question generation tasks relied either on rule-based approaches [19], or complex NLP pipelines [8]. Recent works have started to address this task with recent neural network-based machine learning algorithms [20, 5, 12]. However, some issues still remain. In the majority of the previous works, there are two common challenges: the first is the complexity of analyzing and building a pipeline that is able to generate a natural and relevant question, and the second is data collection. Many researchers utilize crowdsourcing which saves time when it comes to statistical approach, but more expensive. Here we aim to design the experiment that is both efficient and robust in generating natural and relevant questions.

To address the above-mentioned challenges we utilize end-to-end machine learning approaches in order to learn the sentence-question pattern automatically. The utilization of machine learning enables us to do soft-matching [3, 13]. This feature helps with the robustness of the model, enabling it to deal with question patterns that are not available in the training dataset. Recent success of neural machine translation (NMT) technology [2] promises us a high-performance translation result. However, there have also been some reports of traditional statistical machine translation (SMT) approaches that boast better performance, especially in certain low-resource conditions [6]. Inspired by this finding, we utilize these two popular machine translation paradigms in our question generation task.

The power of NMT lies within the attention mechanism and bi-directional architecture properties [9]. The attention mechanism allows us to selectively focus on parts of the source sentence during translation, while the bi-directional architecture enables us to capture information regarding long-term dependency structure of sentences. On the other hand, SMT gains an advantage from phrase-alignment that is statistically calculated during the learning process [7], this alignment is basically a mapping rule that is learned via the training data.

In this work we make the following contributions:

- To our knowledge, we are the first to utilize end-to-end machine learning translation techniques for question generation on Japanese sentences. And we are among the first to employ a deep sequence-to-sequence learning approach to generate questions.
- We propose a simple, but effective way to perform system combination between two popular machine translation paradigm in the Japanese question generation. Experimental results demonstrate that our combined system shows promise for overcoming the shortcomings of each approach.

3 Question Generation

We formulate our question prediction task as follows: Given an input context sentence c, we aim to generate a natural question q. Both input and output can be a sequence of arbitrary length $[c_1, ..., c_{|c|}]$ and $[q_1, ..., q_{|q|}]$.

This question generation task can be defined as:

$$\hat{q} = \arg\max_{q} P(q|c) \tag{1}$$

where $\hat{q}$ is the system best-generated question, and $P(q|c)$ is the conditional probability of the predicted question sequence q, given the input c. Here, we aim to maximize $P(q|c)$ over all possible q. This conditional probability can be likened to a translation model in a statistical sequence learning approach, or a conditional log-likelihood in neural sequence learning approach.

3.1 Rule Based Question Generation

The Japanese language falls under the class of agglutinative languages. Verbal expression is fundamentally located at the end of a sentence, and it can be added to various words such as nouns, particles, and auxiliary verbs with conjugations at the ending of the word stem. If we convert an affirmative sentence to yes-no question form, we need to add an appropriate sentence ending particle. For generating a wh-type question, we need not only that particle, but also need to use an interrogative word.

We analyzed the sentences which end with the following part-of-speech: verbs, adjectives, particles, and auxiliary verbs; we will also consider a few nouns that function similarly to adjectives, verbs, or auxiliary verbs. While creating rules, only “KA(か)” was used as the sentence ending particle in interrogative expressions, because this is the most popular and widely used with any part-of-speech. An interrogative expression and an edit flag (use or delete) of each 3 words are annotated to all sentences by the human annotator. If several sentences have the same interrogative expression and share the same words, they are merged. Finally, we obtained approximately 1,700 converting rules.

To generate wh-type questions, we apply our above-described rules and randomly replace only one word with “how” or “what”. The words which can be replaced are (1) a noun in an objective case, (2) an adjective, or (3) a verb.

3.2 Statistical Question Generation

Here we utilize phrase-based statistical machine translation (SMT) to capture the pattern between context sentence input and question output. We treat the sentence-question pair as a parallel corpus to train the translation model, which is based on the noisy channel model. Considering the general question generation task, we reformulate Equation 1 with Bayes rule as

$$\hat{q} = \arg\max_{q} P(c|q)P_{LM}(q). \tag{2}$$

This way we can obtain the language model $P_{LM}(q)$ and separate the translation model $P(c|q)$.

3.3 Neural Question Generation

Here we generate the question with the neural machine translation (NMT) technique. In contrast with the SMT, the power of NMT lies on the bi-directional recursive architecture and global attention mechanism. The bi-directional architecture enables the model to learn in both a forward and backward context. The attention mechanism allows the model to put emphasis on a certain part of the sentence, imitating the way humans think to solve a task.

Taking into account the question generation task formulation on Equation 1, we can factorize the conditional probability $p(q|c)$ as

$$P(q|c) = \prod_{t=1}^{|q|} P(q_t|c, q_{1..t-1}). \tag{3}$$

Where q_t are word candidates that combine to give output question q, we treat the conditional probability $P(q|c)$ as a product of word-level prediction. The probability of q_t is predicted based on the input context sentence c, and all the words that have been previously generated $q_{1..t-1}$.

Furthermore, we can break down the word-level conditional probability into

$$P(q_t|c, q_{1..t-1}) = softmax(W_q \tanh(W_b[b_t; a_t])). \tag{4}$$

Where b_t portrays the bi-directional recursive network state variable at the time step t, a_t is the attention based encoding of input sentence context c at decoding time step t, and W_q and W_b are the parameters to be learned.

3.4 Vote Mechanism

We devise a simple voting mechanism based on how close the output between the proposed approaches to the question language model in the term of perplexity. The question language model is obtained from the question training dataset. Combination approach (denoted by COM) obtained question with the lowest perplexity from both proposed approaches. More details of this mechanism can be seen in Fig. 1.

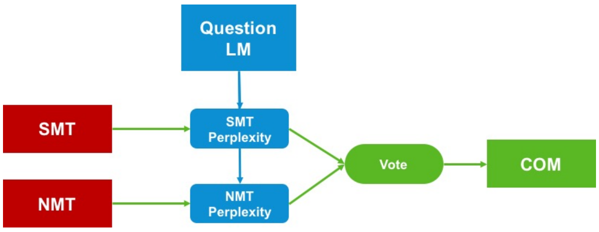

Figure 1: Overview of our question voting mechanism.

4 Datasets

The sentence-question pairs dataset that we constructed in this study is based on the user-merchant review pages [16] on the Rakuten Japan website[1]. We collect the review data from the selected wine genre products from 2013 to 2017. Overall we obtain 241,794 reviews and segment it into 673,963 sentences.

From there, we run our rule-based question generation on all of the review sentences, and employ annotators to check and correct the generated questions. In the end, we managed to gather around 30k data points.

[1] https://www.rakuten.co.jp

	SMT	SMT_CMP	NMT	NMT+PRE	NMT_CMP	NMT_CMP+PRE	COM	RULE
YN	91.70	64.25	38.30	33.38	42.32	42.66	71.19	71.51
WH	75.66	81.75	66.45	69.74	71.70	73.37	81.77	83.53

Table 1: Automatic evaluation result with BLEU score given various approaches.

We split this dataset into two different categories YN and WH, which consecutively portray yes-no and how-what type question, and then further randomly subdivide each of these categories into training, development, and test sets.

	Training	Development	Test	Total
YN	15,755	300	300	16,355
WH	14,571	600	300	15,471

Table 2: Dataset statistics.

However, due to the limited amount of data, we only take 300 pairs of sentence-question as a test set. The results of our experiments will later be presented in the context of this test set.

The details of this dataset are provided in Table 2. There are around 15k training set for each YN and WH category, making total 30k dataset for training. We train the model separately with the different dataset for each YN and WH. Furthermore, we are also employing domain adaptation technique, making a compound model that trained from both YN and WH dataset.

5 Experimental Evaluation

In this paper, we employ two types of approaches: statistical question generation, and neural question generation. We employ Moses[2] toolkit and OpenNMT system [3] for SMT and NMT implementation, respectively.

During SMT training, we employ GaCha filtering [17] to remove noisy sentence level alignment, with the GaCha filtering threshold set to 0.8. In our experiment, the SMT model is denoted by SMT.

As for the NMT training, we employ the Japanese Wikipedia[4] dataset provided by Polyglot Project [1] to enrich the model word embedding. Using this dictionary, we learned the word representation with FastText [4]. The NMT model that utilizes this embedding is indicated by +PRE.

In the NMT-based approach, we follow the same configuration used by Du et al. [5]. Therefore, the model denoted by NMT is treated as the state-of-the-art baseline.

Going a step further, we build a compound model using a combined dataset. This model is denoted by CMP (SMT_CMP, NMT_CMP, and NMT_CMP+PRE). This compound model is trained using a compound dataset (YN and WH). In contrast, the non-compound models (SMT, NMT, and NMT+PRE) are trained with the corresponding training and test sets. For example, in a non-compound model, if we evaluate the model with the YN dataset, we train the model only with the YN dataset. From this compound model, we would like to assess how the model works given a complex and noisy training set.

Not to be confused with the compound model, here we also introduce the voting mechanism that combines both SMT and NMT approaches (denoted by COM). This approach combines the best model on both SMT and NMT setup. Here we utilize SRILM[5] toolkit to help us build the language model.

At the end of this experiment, we evaluated our system response both automatically by calculating the generated response $\hat{q}$ with the question in reference database q, and subjectively by giving out a survey to humans.

5.1 Automatic Evaluation

Automatic evaluation of our models are done by the BLEU-4 metric. This metric calculates how well the generated question compares to the reference question. The results of the BLEU-4 evaluation is presented above in Table 1.

5.2 Subjective Human Evaluation

Next, the human evaluation studies are performed to measure the quality of questions. We conduct the evaluation on rule-based approach (RULE), both SMT approaches (SMT, SMT_CMP), the best model in NMT approaches according to the automatic evaluation results (NMT_CMP+PRE), and the combination vote approach COM. We apply two metrics, content quality and fluency, as defined by the Japanese Patent Office[6]. Content quality indicates grammaticality and consistency with the reference, and the results of this evaluation can be seen in Fig. 2. Fluency focuses on only readability of the generated questions, with the results of this is shown in Fig. 3. If a generated question is neither of interrogative form or question type, conflict happens between YN and WH, and so both scores should be lower.

Here, we randomly sampled 50 sentence-question pairs for YN and WH from these two cases, and evaluated by asking two Japanese native speakers to rate the pairs in terms of the metrics above on a 1-5 scale (5 for the best).

5.3 Discussion

Automatic evaluation results (Table 1) shows that the SMT approach (SMT and SMT_CMP) in general performs better than NMT (NMT and NMT_CMP) and the rule-based RULE approach. Except on the WH type

[2] http://www.statmt.org/moses

[3] http://opennmt.net

[4] http://ja.wikipedia.org

[5] http://www.speech.sri.com/projects/srilm

[6] http://www.jpo.go.jp/shiryou/toushin/chousa/pdf/tokkyohonyaku_hyouka/01.pdf

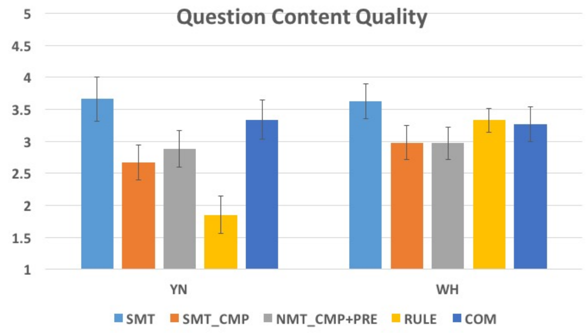

Figure 2: Subjective evaluation result on content quality.

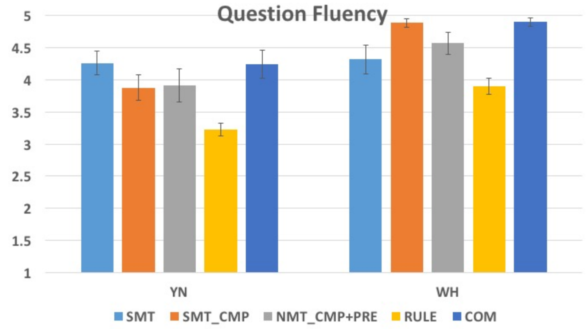

Figure 3: Subjective evaluation result on question fluency.

dataset, where the rule-based approach performs a little bit better. However, during the subjective evaluation, we found out that the content quality and fluency of the SMT approach is much better. The machine learning's soft-matching feature allows the model to become more flexible in generating natural questions, while the rule-based generated question looks like a canned question.

Overall the combination vote system not showing significant improvement in BLEU metrics. However, on both quality and fluency subjective evaluation, we can see that COM approach perform on par and/or significantly better compared to others. It indicates that the voting mechanism manages to select a more natural and comprehensible question.

6 Conclusion

A data-driven automatic question generation approach for Japanese text is presented. Experimental evaluation shows that the statistical question generation model achieves the best performance on Japanese text in both automatic and subjective evaluation. Furthermore, a combination system that votes the output between SMT and NMT is also introduced. This combination system performs better in both subjective evaluation metrics, overcoming the shortcomings of each approach. As future work, implementing this question generation to the real FAQ generation problem can be a promising future direction.

References

[1] Al-Rfou, R., Perozzi, B., and Skiena, S. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the 17th Conference on Computational Natural Language Learning* (2013), CoNLL '13, Association for Computational Linguistics, pp. 183–192.

[2] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations* (2015), ICLR '15.

[3] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. A neural probabilistic language model. *The Journal of Machine Learning Research 3* (2003), 1137–1155.

[4] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics 5* (2017), 135–146.

[5] Du, X., Shao, J., and Cardie, C. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (2017), ACL '17, Association for Computational Linguistics, pp. 1342–1352.

[6] Koehn, P., and Knowles, R. Six challenges for neural machine translation. In *Proceedings of the 1st Workshop on Neural Machine Translation* (2017), Association for Computational Linguistics, pp. 28–39.

[7] Koehn, P., Och, F. J., and Marcu, D. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (2003), NAACL '03, Association for Computational Linguistics, pp. 48–54.

[8] Labutov, I., Basu, S., and Vanderwende, L. Deep questions without deep understanding. In *Proceedings of the 7th International Joint Conference on Natural Language Processing* (July 2015), IJCNLP '15, Association for Computational Linguistics, pp. 889–898.

[9] Luong, T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015), Association for Computational Linguistics, pp. 1412–1421.

[10] Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., and Vanderwende, L. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016), ACL '16, Association for Computer Linguistics, pp. 1802–1813.

[11] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. Ms marco: A human generated MAchine Reading COmprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[12] Nio, L., and Murakami, K. Intelligence is asking the right question: A study on japanese question generation. In *Proceedings of the IEEE Spoken Language Technology Workshop* (2018).

[13] Nio, L., Sakti, S., Neubig, G., Yoshino, K., and Nakamura, S. Neural network approaches to dialog response retrieval and generation. *IEICE Transactions 99-D*, 10 (2016), 2508–2517.

[14] Piwek, P., Hernault, H., Prendinger, H., and Ishizuka, M. T2D: Generating dialogues between virtual agents automatically from text. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (2007), IVA '07, Springer Berlin Heidelberg, pp. 161–174.

[15] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016), EMNLP '16, Association for Computational Linguistics, pp. 2383–2392.

[16] Shinzato, K., and Oyamada, Y. What do people write in reviews for sellers? investigation and development of an automatic classification system. *Journal of Natural Language Processing 25*, 1 (2018).

[17] Tan, L., and Pal, S. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation* (2014), Association for Computational Linguistics, pp. 201–206.

[18] Tang, D., Duan, N., Qin, T., and Zhou, M. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027* (2017).

[19] Y., C., and Hasan, S. A. Towards automatic topical question generation. In *Proceedings of the 24th International Conference on Computational Linguistics* (2012), COLING '12, Association for Computational Linguistics, pp. 475–492.

[20] Yang, Z., Hu, J., Salakhutdinov, R., and Cohen, W. Semi-supervised qa with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (2017), ACL '17, Association for Computational Linguistics, pp. 1040–1050.