Distant Supervision for Extractive Question Summarization

Tatsuya Ishigaki[1], Kazuya Machida[1], Hayato Kobayashi[2], Hiroya Takamura[1,3], and Manabu Okumura[1]

[1] Tokyo Institute of Technology,
{ishigaki, machida}@lr.pi.titech.ac.jp, oku@pi.titech.ac.jp
[2] Yahoo Japan Corporation / RIKEN AIP, hakobaya@yahoo-corp.jp
[3] AIST, takamura.hiroya@aist.go.jp

Abstract. Questions are often lengthy and difficult to understand because they tend to contain peripheral information. Previous work relies on costly human-annotated data or question-title pairs. In this work, we propose a distant supervision framework that can train a question summarizer without annotation costs or question-title pairs, where sentences are automatically annotated by means of heuristic rules. The key idea is that a single-sentence question tends to have a summary-like property. We empirically show that our models trained on the framework perform competitively with respect to supervised models without the requirement of a costly human-annotated dataset.

Keywords: Question Summarization · Extractive Summarization.

1 Introduction

People ask questions in various scenarios such as community question answering (CQA) sites, conference sessions, and e-mail conversations. However, questions tend to be lengthy and laborious to understand because they often contain peripheral information. Question summarization is the task of transforming such questions into shorter and concise ones. We focus on the setting used in existing studies that transforms an input question into a concise single-sentence summary [21, 12, 7]. This setting can be regarded as title or headline generation, and is particularly important for practical application on CQA sites where questions do not always have appropriate titles, unlike news articles. In fact, to reduce the burden on users who post questions, many CQA sites (including the biggest Japanese CQA site, Yahoo! Chiebukuro [23]), do not provide an input field for titles in the submission form to reduce the burden on users who post questions. On most CQA sites, the first sentence of a question is displayed as the headline, but this is not always appropriate. We show these examples in Table 1.

There are two approaches for summarization tasks: extractive and abstractive. The extractive approach selects salient units (e.g., sentences) in the input [15, 8, 16, 20, 13, 3], while the abstractive approach generates a summary by possibly paraphrasing the information in the input [2, 22]. The extractive approach

Gold label	Salient Score	Input question
0	0.17	I am an aged person.
0	0.45	Please kindly tell me the answer in detail.
...	...	...
1	0.99	**How can I write an email with a clickable url?**
0	0.05	I do stretches to get flexibility in the legs.
1	0.95	**How am I able to do a front split?**
...	...	...
0	0.05	All of my family except me can do it.
0	0.96	Is there any reason why only I cannot do a front split?

Table 1: Two translated examples of input question, gold labels (sentences labeled 1 should be included in the summary), and scores for the sentences given by our model `DistNet`. **Bold** and underlined sentences refer to ones selected by our model (`DistNet` +`Init`) and a baseline (`Lead` +`Q`), respectively.

has an advantage in that it does not generate errors when extracting one sentence for title generation and can be directly used for real services, as reported by Higurashi et al. [7]. Therefore, we take the extractive approach assuming that extracted titles will be used for a real service.

Previous studies on question summarization rely on costly human-annotated data or automatically collected question-title pairs. Tamura et al. [21] used an SVM-based classifier to extract the most important sentences to improve the performance of question answering systems. Higurashi et al. [7] proposed a 'learning to rank' approach for headline generation for CQA. Both of these methods require costly human-annotated data. Ishigaki et al. [12] trained extractive and abstractive summarizers by regarding a question-title pair posted on a CQA site as a question-summary pair. However, questions do not always have appropriate titles on other CQA sites, as mentioned above.

To overcome this problem, we take a distant supervision approach instead of creating costly human-annotated data. Distant supervision [17] involves training a model on pseudo data automatically labeled on the basis of heuristics, rules, and/or external resources. Various approaches for different tasks have been proposed, such as the use of FreeBase [1] for relation extraction, emoticons for sentiment classification [6], and topic labels on Wikipedia articles for blog topic classification [10]. In the summarization field, Nallapati et al. [18] and Chen and Lapata [3] automatically annotated labels for sentence extraction by using a dataset with human-generated abstractive summaries. However, we face difficulties in directly applying their strategies due to the lack of human-generated abstractive summaries (or titles).

The key idea in this paper is to focus on the difference in characteristics between single-sentence questions and individual sentences in extremely long multi-sentence questions posted on CQA sites. We assume that the single-sentence questions are mostly self-contained questions, and can function as a summary (or title), while the individual sentences in long multi-sentence questions are not. Thus, we use a classifier to determine how likely it is that a sentence is similar to a single-sentence question and would therefore be appropriate as a title.

Our contributions are as follows:

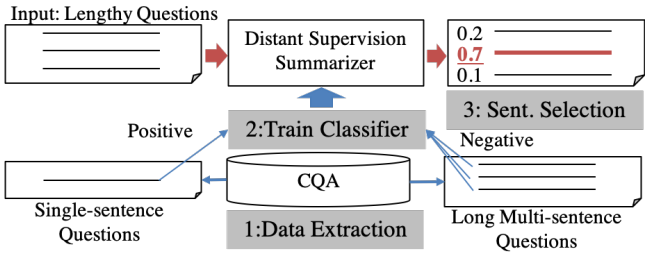

Fig. 1: Overview of our framework.

- We propose a distant supervision framework for question summarization on CQA sites and verify the correctness of the assumption through manual analysis.
- We construct a large training dataset including 2.5M sentences with pseudo labels from Yahoo! Chiebukuro and will release it [11].
- We compare our models with several baselines on a human-annotated evaluation dataset and find that our models perform competitively with respect to the supervised baselines.

2 Framework

Correctness of Assumption: The ideal summaries for questions should be 1) a question sentence and 2) self-contained. We manually analyzed 300 randomly selected single-sentence questions on Yahoo! Chiebukuro and found that 98% of them were self-contained questions. In contrast, 91.2% of 315 sentences randomly selected from long multi-sentence posts consisting of more than ten sentences were not self-contained questions. In fact, they are often not even a question, which makes them more similar to sentences not to be included in a summary than to titles. Our main finding here is that single-sentence questions on CQA sites are mostly self-contained, which makes them similar to sentences used as a title. Note that small amounts of noise or incorrect labels can be ignored, as shown in later in the experimental results (Sec. 3).

Overview of Proposed Framework: Figure 1 shows the overview of our proposed framework, which includes the following three steps:

1. Extract single-sentence questions and extremely long multi-sentence questions from a CQA site, whose sentences can be automatically labeled by the assumption.
2. Train a classifier (that can predict a label with its confidence score) by using the automatically annotated sentences as pseudo training data.
3. Use confidence scores determined by the classifier to select a sentence from an input question document to form a summary.

In Step 1, we can obtain a large number of positive and negative instances without annotation costs. For positive instances, 28% of the questions on Yahoo! Chiebukuro are single-sentence questions. For negative instances, only a few

long multi-sentence questions exist, but the number of sentences in the questions is large. In Step 2, we can use any classifier even with a large number of parameters thanks to the availability of large-scale data. In our experiments, we examine neural network and logistic regression classifiers. In Step 3, we prepare three selection strategies: `Greedy`, `Init`, and `Q`. `Greedy` selects the sentence with the highest confidence score, whereas `Init` selects the initial sentence among sentences with reasonably high confidence scores, which is expected to be effective for multi-focused questions. `Q` first extracts question sentences by means of a rule-based approach and then selects the sentence with the highest score among the extracted question sentences. We will explain these steps (i.e., how to create the pseudo data, the training settings of the classifiers, and the algorithms for the selection strategies) in more detail in the next section.

3 Experiments

Data: We created two datasets, `Pseudo` and `Label`, from the publicly available dataset [19] provided by Yahoo! Chiebukuro. `Pseudo` is a large dataset with pseudo labels for training classifiers. `Label` is a dataset with human-annotated correct answers.

`Pseudo` contains 800K single-sentence questions as positive instances and 1.7M sentences extracted from multi-sentence questions as negative instances. We randomly extracted single-sentence questions and multi-sentence questions consisting of more than ten sentences and assigned positive/negative labels on the basis of the assumption. Note that classifiers trained with `Pseudo` can accept any sentence even in a question with more than ten sentences since they do not care about the question length.

`Label` consists of 12,406 questions (including multiple sentences) separately sampled from `Pseudo`. Every sentence in each question has a binary label indicating whether the sentence is the most important, that is, only the best sentence has a label of 1, and the others have 0. We used crowdsourcing to annotate `Label`, where given a question, five workers were asked to select the best sentence that represented the main focus of the question. We included only questions for which at least four workers selected the same sentence.

Compared Models: We compared two models on our framework with several unsupervised and supervised baselines as follows. `DistNet` and `DistReg` are variants based on neural network and logistic regression models, respectively.

- `DistNet` selects the most important sentence on our framework with a neural network classifier trained on `Pseudo`, which uses an LSTM [9] to encode the input question into a fixed-length vector, and a softmax layer to convert the vector into a probability distribution expressing how likely it is that the sentence is similar to a single-sentence question.
- `DistReg` is a variant based on a logistic regression classifier [4] with n-gram, part-of-speech features.

The unsupervised baselines are as follows.

- `Lead` selects the first sentence, which is strong for generic summarization.

- `LexRank` selects the most important sentence using a graph-based extractive summarization method [5] trained on all the questions in the dataset.
- `SimEmb` selects the sentence that is the most similar to the input question in terms of the word mover's distance [14], where word embeddings were trained on all the questions. Taking account of the similarity between the input document and a sentence is common in unsupervised models for generic summarization.
- `TfIdf` selects the sentence with the highest averaged TF-IDF score, where the IDF was calculated by using all the questions.

The supervised baselines are variants of `DistNet` and `DistReg`.

- `SupNet` selects the most important sentence based on a neural network classifier trained on `Label` in the same way as `DistNet`.
- `SupReg` is a variant of `DistReg` based on a logistic regression classifier trained on `Label`.

Sentence Selection Strategies: We used three sentence selection strategies `Greedy`, `Init` and `Q`, as in Sec. 2.

- `Greedy` selects the sentence with the highest score.
- `Init` selects the initial sentence that is given a higher score than the specific threshold τ. τ was tuned on the validation data (as described later).
- `Q` selects the best one among the question sentences determined by the rule-based question extractor [21]. If there are no question sentences, `Q` is the same as `Greedy`.

Evaluation: For evaluating the performance, we used the accuracy measure calculated by dividing the number of questions in which the target method correctly selected their most important sentences by the number of questions used. Note that well-known metrics such as ROUGE and precision/recall are not appropriate since our task is to find only one sentence as a title (or snippet for a list item). We divided the labeled data `Label` into five sets (training:development:test=3:1:1) and performed 5-fold cross-validation for evaluating the supervised models and tuning τ.

Results: Table 2 shows the accuracy scores for the compared models with three strategies. The numbers in bold represent the best performing models that are trained without labeled data. There is no statistically significant difference between them and `SupReg` even though they do not use labeled data.

	Greedy	Init	Q	Best
DistNet	**87.38**	**90.45**	**87.38**	**90.45**
DistReg	86.17	89.05	86.17	89.05
Lead	81.79	81.79	88.08	88.08
LexRank	78.49	81.79	84.95	84.95
SimEmb	59.46	81.79	71.17	81.79
TfIdf	52.03	81.79	69.68	81.79
SupNet	81.67	86.31	81.67	86.31
SupReg	87.89	91.21	87.89	91.21

Table 2: Accuracy for compared models

Looking at the column of `Greedy`, our model `DistNet` performed competitively with the best supervised model `SupReg`, although `DistNet` did not use the labeled dataset `Label`. Note that we didn't observe any statistically significant difference between `DistNet` and `SupReg` by the sign test. Comparing `DistReg` and `SupReg`, `SupReg` conversely outperformed `DistReg` because simple models like logistic regression can be trained on a small dataset. This implies

that our distant supervision approach is suitable for complicated models like LSTM. Among unsupervised baselines, `Lead` performed the best, and the others did not perform well. This may be because `LexRank`, `SimEmb`, and `TfIdf` do not take into account whether a sentence is a question.

The second column `Init` shows the results of selecting the initial sentence among the sentences with reasonably high scores. Comparing `Greedy` and `Init` for `DistNet`, we found that the simple strategy `Init` drastically improved the accuracy. This matches our intuition that former sentences tend to be important. Note that `Lead` with `Init` is the same as `Lead` with `Greedy`. `LexRank`, `SimEmb`, `TfIdf` with `Init` became equivalent to `Lead` as a result of tuning.

The third column `Q` shows the results of selecting the best sentence among the question sentences. This column is mainly for unsupervised baselines. The results of our models and supervised baselines are the same as `Greedy` since they were trained for basically extracting question sentences. The results in the column indicate that, overall, their accuracy scores were further enhanced, but as a result, none of the unsupervised baselines could overtake our best model `DistNet` with `Init`. We also examined the unsupervised baselines with both `Init` and `Q`, but the trends did not change.

The final column shows the best scores in the left three columns. Our model `DistNet` with `Init` outperformed the best all the baselines trained without `Label`. Furthermore, `DistNet` performed competitively with the best performing supervised model `SupReg` with `Init`, since we didn't observe any statistical significance between them.

Qualitative analysis: Table 1 shows two examples of our model `DistNet` with `Init` (bold) and the strong baseline `Lead` with `Q` (underlined), where the scores were calculated by `DistNet`. The first example shows that our model successfully selected the self-contained question, while `Lead` with `Q` failed. The rule-based question extractor sometimes failed because it possibly selects non self-contained questions such as the second sentence in this example. The second example shows how effective `Init` was. We observed that `DistNet` with `Greedy` sometimes failed when the input was multi-focused, as shown in the last sentence (0.96), since the `DistNet` gave very high probabilities for question sentences. In contrast, `Init` mostly handled multi-focused questions well, thereby delivering better performances in terms of accuracy.

4 Conclusion

We proposed a distant supervision framework for question summarization based on the assumption that single sentence posts tend to have a summary-like property. For future research, we will examine if our assumption can be applied to other domains such as answers on CQA sites, user comments on news sites, and opinions on discussion forums.

References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD2008). pp. 1247–1250 (2008)
2. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H.: Distraction-based Neural Networks for Modeling Documents. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI2016). pp. 2754–2760 (2016)
3. Cheng, J., Lapata, M.: Neural Summarization by Extracting Sentences and Words. In: Proceedings of 2016 Annual Conference of the Association for Computational Linguistics (ACL2016). pp. 484–494 (2016)
4. Cox, D.R.: The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. Series B (Methodological) pp. 215–242 (1958)
5. Erkan, G., Radev, D.R.: Lexrank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research **22**, 457–479 (2004)
6. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision. In: Stanford Technical Report (2009)
7. Higurashi, T., Kobayashi, H., Masuyama, T., Murao, K.: Extractive Headline Generation Based on Learning to Rank for Community Question Answering. In: Proceedings of The 27th International Conference on Computational Linguistics (COLING2018). pp. 1742–1753 (2018)
8. Hirao, T., Isozaki, H., Maeda, E., Matsumoto, Y.: Extracting Important Sentences with Support Vector Machines. In: Proceedings of The 11st International Conference on Computational Linguistics (COLING2002). pp. 1–7 (2002)
9. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997)
10. Husby, S.D., Barbosa, D.: Topic Classification of Blog Posts using Distant Supervision. In: Proceedings of the Workshop on Semantic Analysis in Social Media. pp. 28–36 (2012)
11. Ishigaki, T.: Scripts for Preprocessing Yahoo Chiebukuro dataset. http://lr-www.pi.titech.ac.jp/~ishigaki/chiebukuro/ (2020)
12. Ishigaki, T., Takamura, H., Okumura, M.: Summarizing Lengthy Questions. In: Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP2017). vol. 1, pp. 792–800 (2017)
13. Kågebäck, M., Mogren, O., Tahmasebi, N., Dubhashi, D.: Extractive summarization using continuous vector space models. In: Proceedings of The 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC2014). pp. 31–39 (2014)
14. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From Word Embeddings to Document Distances. In: Proceedings of The 32nd International Conference on Machine Learning (ICML2015). pp. 957–966 (2015)
15. Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
16. Mihalcea, R., Tarau, P.: TextRank: Bridging Order into Texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004). pp. 404–411 (2004)
17. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP2009). pp. 1003–1011 (2009)

18. Nallapati, R., Zhai, F., Zhou, B.: SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In: Proceedings of Thirty-First AAAI Conference on Artificial Intelligence (AAAI2017). pp. 3075–3081 (2017)
19. NII: Yahoo! Chiebukuro data (2nd edition). https://www.nii.ac.jp/dsc/idr/en/yahoo/ (2018)
20. Takamura, H., Okumura, M.: Text Summarization Model based on Maximum Coverage Problem and its Variant. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). pp. 781–789 (2009)
21. Tamura, A., Takamura, H., Okumura, M.: Classification of Multiple-Sentence Questions. In: Proceedings of The 2nd International Joint Conference on Natural Language Processing (IJCNLP2005). pp. 426–437 (2005)
22. Tan, J., Wan, X., Xiao, J.: Abstractive Document Summarization with a Graph-based Attentional Neural Model. In: Proceedings of The 55th annual meeting of the Association for Computational Linguistics (ACL2017). vol. 1, pp. 1171–1181 (2017)
23. Yahoo Japan Corp.: Yahoo! Chiebukuro. https://chiebukuro.yahoo.co.jp/ (2019)