Evaluating a How-to Tip Machine Comprehension Model with QA Examples collected from a Community QA Site

Tingxuan Li, Shuting Bai, Fuzhu Zhu, Takehito Utsuro
Degree Programs in Systems and Information Engineering,
Graduate School of Science and Technology, University of Tsukuba,
1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan

Abstract

In the field of factoid question answering (QA), it is known that the state-of-the-art technology has achieved an accuracy comparable to human. However, in the area of non-factoid QA, there are only limited numbers of datasets for training QA models. So within the field of the non-factoid QA, Chen et al. (2020) developed a dataset for training Japanese tip QA models. Although it can be shown that the trained Japanese tip QA model outperforms the factoid QA model, this paper further aims at answering tip questions more closely related to daily lives. Specifically, we collect community QA examples from a community QA site and then apply the trained Japanese tip QA model to those community QA examples. Evaluation results show that the trained tip QA model outperforms the factoid QA model when testing against those community QA examples.

1 Introduction

Among factoid QA related to Wikipedia articles and news articles, the Stanford Question Answering Dataset (SQuAD) (Pranav et al., 2016; Pranav et al., 2018) is one of the most well-known QA datasets and benchmark tests. SQuAD is a reading comprehension dataset. It consists of questions posed by crowdworkers on a set of Wikipedia articles. The answer to every question is a text segment, or span, from the corresponding reading passage, or the question might be unanswerable. The state-of-the-art machine comprehension models trained with SQuAD outperform humans (Devlin et al., 2019; Zhang et al., 2019). However, there is a relatively limited number of published literature that handles the development of datasets for non-factoid QA and the application of state-of-the-art general-purpose machine comprehension models to those non-factoid datasets. Typical non-factoid QA tasks include opinion QA, definition QA, reason QA, and how-to tip QA.

Among various kinds of non-factoid knowledge which are the key to developing techniques for non-factoid QA tasks, Chen et al. (2020) studied how to develop a dataset for training Japanese how-to tip (following Chen et al. (2020), we use the simplified term "tip") QA models. As examples in Chen et al. (2020), tip QA datasets for 'job hunting," "marriage," "apartment," "hay fever," "dentist," and "food poisoning" are developed. "Job hunting" and "marriage" tip QAs are for both training and testing, while other tip QAs are only for testing. For "job hunting", Figure 1 presents a typical example of a tuple of a context, a tip question, and an answer[1].

Chen et al. (2020) applied BERT (Devlin et al., 2019), one of the state-of-the-art machine comprehension models, to a Japanese tip QA dataset. The trained tip QA model is also compared with a factoid QA model which is also trained with a Japanese factoid QA dataset. Evaluation results revealed that the tip machine comprehension performance was al-

[1]The example is extracted from a column web page entitled "Formatting Tips for Your Curriculum Vitae (CV)" (`https://www.thebalancecareers.com/curriculum-vitae-format-2060351`) from a tip website titled "The Balance Careers" (`https://www.thebalancecareers.com/`).

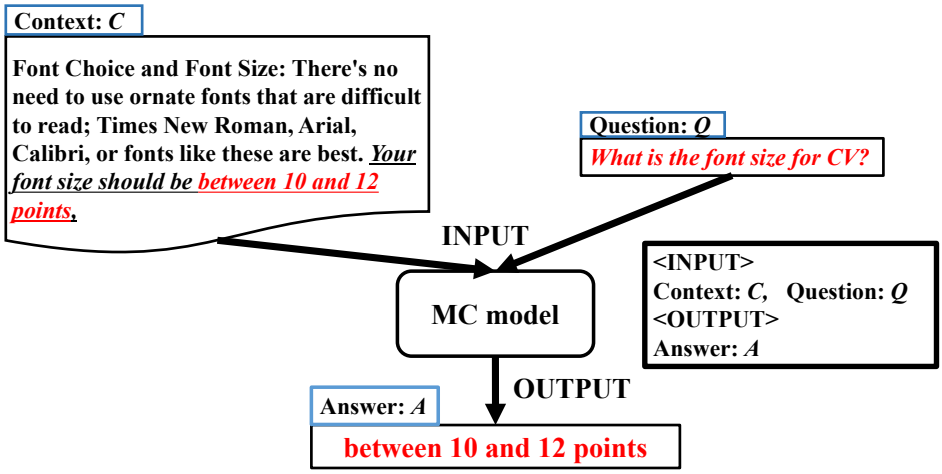

Figure 1: An example of the machine comprehension model of tip QA for "job hunting" together with an example of a tuple of a context C, a question Q, and an answer A

most comparative with that of the factoid machine comprehension even with the training data size reduced to around 4% of the factoid machine comprehension. Thus, the tip machine comprehension task requires much less training data compared with the factoid machine comprehension task.

Following those discussion above, this paper focuses on a Community QA site (for example, "Yahoo! Chiebukuro"[2]), that is different from tip websites and has many pairs of question and answer about the issues people frequently encounter in their real lives. This paper specifically collects tip QAs about "job-hunting" from "Yahoo! Chiebukuro". Compared with the tip QAs collected from tip websites related to job-hunting, these tip QAs include questions posted by people who ask about their confuse based on their experience in the real lives. Moreover, answers are written by veterans and specialists. Therefore, this paper regards the question and answer as the question Q and the context C in a tip QA respectively so that the collected tip QAs can be used as the testing dataset for the machine comprehension model. Evaluation result shows that the performance of the tip machine comprehension model trained in section 2 outperforms the factoid machine comprehension model.

2 Training Tip Machine Comprehension Model

In the training of the tip machine comprehension model, we follow Chen et al. (2020) for the details of the training. We focus on evaluation results with the SQuAD2.0 type dataset. Roughly speaking, for the factoid QAs, the training dataset consists of about 28,000 answerable and 29,000 unanswerable QA examples. For the tip QAs, on the other hand, the training dataset consists of about 1,100 answerable and 12,000 unanswerable QA examples.

As the version of BERT (Devlin et al., 2019) implementation which can handle a text in Japanese, the TensorFlow version[3] and the Multilingual Cased model[4] were used as the pre-trained model. The BERT fine-tuning module for machine comprehension was applied as well as the fine-tuned model. The BERT pre-trained model was fine-tuned with

[2] https://chiebukuro.yahoo.co.jp

[3] https://github.com/google-research/bert

[4] Trained with 104 languages, available from https://github.com/google-research/bert/blob/master/multilingual.md.

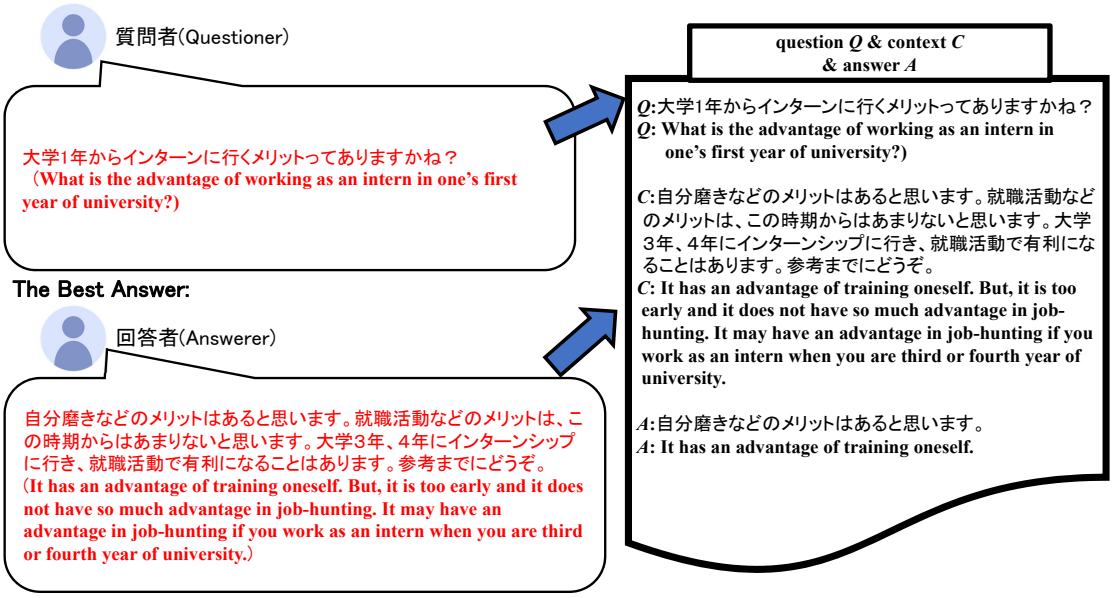

Figure 2: An Example of Collecting Community QA Examples from a Community QA Site ("Yahoo! Chiebukuro")

the following three types of training datasets:

(i) The training dataset of factoid QAs.

(ii) The training datasets of the tip QA about "job hunting" and "marriage".

(iii) Mix of (i) and (ii).

Note that we train a single model with each of these three training datasets (i)∼(iii).

3 Developing a How-to Tip Machine Comprehension Dataset from a Community QA Site

This section describes the procedure (Figure 2) for developing a tip machine comprehension dataset from a community QA site. About the community QA site, we choose "Yahoo! Chiebukuro"[5] that is maintained by Yahoo! JAPAN. "Yahoo!Chiebukuro" is an online knowledge community where people can share and communicate about tip and knowledge in real lives. About "Yahoo! Chiebukuro", we list the format of the parts that we use in this paper in the following.

Category Name: The medium or small category that appears at the end of the "Category Path" (for example: "job hunting").

Category Path: Large, medium, small categories connected by '>' (for example: > job and career > job change and job hunting > job hunting).

Question Title: More than 5 Japanese characters and less than 100 Japanese characters.

Question Content: Less than 1,000 Japanese characters.

Answer Content: The complete content of the answer. More than 10 Japanese characters and less than 2,000 Japanese characters.

In this section, following the format above, we collect the question Q, the context C, and the answer A of tip QAs manually from "Yahoo! Chiebukuro" and use them as the testing dataset of tip machine comprehension model. As for the evaluation, it is manually judged whether the answer $\hat{A}$ predicted by the fine-tuned model matches the real answer A or not. Specifically, we set the "Category Name" and

[5]This paper uses the third version of "Yahoo! Chiebukuro" (provided in 2020 that includes the data from 2015/04/01 to 2018/03/31.

Table 1: Statistics of Testing Dataset collected from "Yahoo! Chiebukuro"

keyword	# words within a context (simplified/not simplified)		# words of a question (simplified/not simplified)	
	shortest∼longest	average	shortest∼longest	average
interview	$34 \sim 303/34 \sim 419$	$118.1/126.2$	$12 \sim 60/15 \sim 364$	$43.7/117.5$
clothe, hairstyle	$34 \sim 92/34 \sim 209$	$64.3/83.1$	$12 \sim 44/26 \sim 211$	$26.7/85.8$
letter, internship	$13 \sim 90/13 \sim 295$	$54.0/93.1$	$8 \sim 42/8 \sim 238$	$24.4/93.2$
total	$13 \sim 303/13 \sim 419$	$94.5/110.9$	$8 \sim 60/8 \sim 364$	$36.4/106.3$

"Category Path" as "job hunting" and ">job and career>job change and job hunting>job hunting", respectively. Besides, we choose several keywords related to job-hunting. Whether the keyword is present in the "Question Title" or the "Question Content" is set as the search condition. The pairs of "Question Content" and "Answer Content" collected according to the above conditions are judged whether they are suitable for the tip QAs used as the testing dataset of tip machine comprehension model.

We evaluate the tip machine comprehension model with the following four datasets.

- Question and context are both simplified.
- Only context is simplified.
- Only question is simplified.
- Neither question nor context is simplified.

According to the following procedure, we develop the dataset of "Simplified Question" and "Simplified Context" and use it for the evaluation of the tip machine comprehension model.

Simplified Question We extract one sentence from "Question Content" as the question Q. If there are multiple questions in the content, we extract the first one.

Simplified Context We extract one paragraph from "Answer Content".

An example of collecting the community QA examples collected from "Yahoo! Chiebukuro" is shown in Figure 2.

We obtain 20 pairs of tip QAs out of the 143 examples collected by searching with the keyword "letter, internship", 20 pairs out of 91 examples collected by searching with the keyword "hairstyle, clothes", and 60 pairs out of 173 examples collected by searching with the keyword "interview". With the total 100 pairs of the tip QAs collected from "Yahoo! Chiebukuro", we develop 100 pairs each of "Question and context are both simplified" tip QAs, "Only context is simplified" tip QAs, "Only question is simplified" tip QAs, and "Neither question nor context is simplified" tip QAs, respectively.

When developing a community QA examples dataset from "Yahoo! Chiebukuro", a total of 407 pairs of the examples are examined. Among those examples, those which have the characteristics such as less relevant to "job hunting", an ambiguous question or answer, short answer, multiple-choice question, negative answer, question including pictures, etc. have been excluded. After excluding 307 examples, remaining 100 examples are considered as with "Answerable Context", indicating that the answer of the question Q is definitely included in the context C. Statistics such as the number of words within the community QA examples collected from "Yahoo! Chiebukuro" is shown in Table 1.

4 Evaluation

4.1 Evaluation Procedure

We apply the tip machine comprehension model to the dataset developed from "Yahoo! Chiebukuro" according to the procedure described in the previous section. The dataset includes 100 pairs each of "Question and context are both simplified", "Only context is simplified", "Only question is simpli-

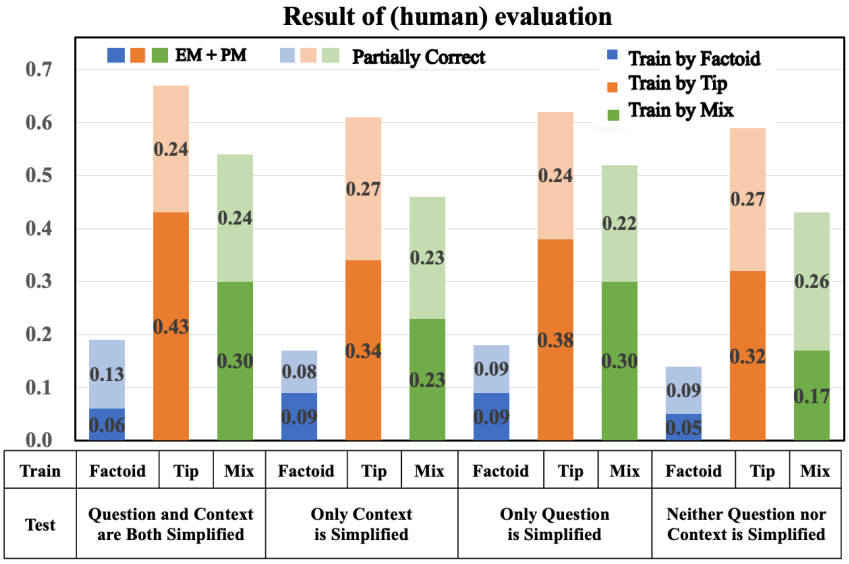

Figure 3: Evaluation Results for Community QA Examples collected from the Community QA Site ("Yahoo! Chiebukuro") (exact match + partial match (EM+PM) with correct answer and partially correct)

fied", and "Neither question nor context is simplified" community QA examples. We manually evaluate the predicted answer according to the 3 levels of exact match + partial match (EM+PM) with correct answer, partially correct, and not correct.

4.2 Evaluation Result

The model fine-tuned with the dataset of the tip QAs has the best performance, outperforming the model fine-tuned with the dataset of factoid QAs and the mixture of datasets of factoid QAs and the tip QAs. Based on the result, the tip machine comprehension model fine-tuned with the dataset of the tip QAs achieves a certain level of performance when applied to the community QA examples collected from "Yahoo! Chiebukuro".

5 Related Work

A limited number of QA datasets including non-factoid QAs are known in any language. In English, MS MARCO (Nguyen et al., 2016) has been developed using Bing's search logs and passages of retrieved web pages. MS MARCO may include non-factoid QAs. In Chinese, DuReader (He et al., 2018) has been developed using Baidu Search and Baidu Zhidao, which is a Chinese community-based QA site. Its QAs include non-factoid ones, where its question types are classified into *entity, description*, and *yes-no* questions on *fact* or *opinion*. NarrativeQA (Kočiský et al., 2018) dataset (in English), which is also a non-factoid type QA dataset, contains questions created by editors based on summaries of movie scripts and books. In the case of the Japanese language QA dataset, quite a limited number of publicly available factoid QA datasets exist, where no Japanese non-factoid QA dataset is publicly available.

6 Conclusion

Chen et al. (2020) developed a dataset for training Japanese tip QA models. Although it can be shown that the trained Japanese tip QA model outperforms the factoid QA model, this paper further aims at answering tip questions more closely related to daily lives. Specifically, we collect community QA examples from a community QA site and then apply the trained Japanese tip QA model to those community QA examples. Evaluation results show that the

trained tip QA model outperforms the factoid QA model when testing against those community QA examples.

Acknowledgements

In this paper, we used "Yahoo! Chiebukuro data (3rd edition)" provided by Yahoo Japan Corporation via IDR Dataset Service of National Institute of Informatics.

References

T. Chen, H. Li, M. Kasamatsu, T. Utsuro, and Y. Kawada. 2020. Developing a how-to tip machine comprehension dataset and its evaluation in machine comprehension by BERT. In *Proc. 3rd FEVER*, pages 26–35.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186.

W. He, K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Liu, Y. Wang, H. Wu, Q. She, X. Liu, T. Wu, and H. Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proc. MRQA*, pages 37–46.

T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

R. Pranav, Z. Jian, L. Konstantin, and L. Percy. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. EMNLP*, pages 2383–2392.

R. Pranav, J. Robin, and L. Percy. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proc. 56th ACL*, pages 784–789.

Z. Zhang, Y. Wu, J. Zhou, S. Duan, and H. Zhao. 2019. SG-Net: Syntax-guided machine reading comprehension. *CoRR*, abs/1908.05147.