Machine Comprehension Improves Domain-Specific Japanese Predicate-Argument Structure Analysis

Norio Takahashi **Tomohide Shibata**[*] **Daisuke Kawahara** **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{ntakahashi, shibata, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

To improve the accuracy of predicate-argument structure (PAS) analysis, large-scale training data and knowledge for PAS analysis are indispensable. We focus on a specific domain, specifically Japanese blogs on driving, and construct two wide-coverage datasets as a form of QA using crowdsourcing: a PAS-QA dataset and a reading comprehension QA (RC-QA) dataset. We train a machine comprehension (MC) model based on these datasets to perform PAS analysis. Our experiments show that a stepwise training method is the most effective, which pre-trains an MC model based on the RC-QA dataset to acquire domain knowledge and then fine-tunes based on the PAS-QA dataset.

1 Introduction

To understand the meaning of a sentence or a text, it is essential to analyze relations between a predicate and its arguments. Such analysis is called semantic role labeling (SRL) or predicate-argument structure (PAS) analysis. For English, the accuracy of SRL has reached approximately 80%-90% (Ouchi et al., 2018; He et al., 2018; Strubell et al., 2018; Tan et al., 2018). However, there are many omissions of arguments in Japanese, and the accuracy of Japanese PAS analysis on omitted arguments is still around 50%-60% (Shibata et al., 2016; Shibata and Kurohashi, 2018; Kurita et al., 2018; Ouchi et al., 2017). A reason for such low accuracy is the shortage of gold datasets and knowledge about PAS analysis, which require a prohibitive cost of creation (Iida et al., 2007; Kawahara et al., 2002).

From the viewpoint of text understanding, machine comprehension (MC) has been actively studied in recent years. In MC studies, QA datasets consisting of triplets of a document, a question and its answer are constructed, and an MC model is trained using these datasets (e.g., Rajpurkar et al. (2016) and Trischler et al. (2017)). MC has made remarkable progress in the last couple of years, and MC models have even exceeded human accuracy in some datasets (Devlin et al., 2019). However, MC accuracy is not necessarily high for documents that contain anaphoric phenomena and those that need external knowledge or inference (Mihaylov et al., 2018; Yang et al., 2018).

In this paper, we propose a Japanese PAS analysis method based on the MC framework for a specific domain. In particular, we focus on a challenging task of finding an antecedent of a zero pronoun within PAS analysis. We construct a wide-coverage QA dataset for PAS analysis (PAS-QA) in the domain and feed it to an MC model to perform PAS analysis. We also construct a QA dataset for reading comprehension (RC-QA) in the same domain and jointly use the two datasets in the MC model to improve PAS analysis.

We consider the domain of blogs on driving because of the following two reasons. Firstly, we can construct high-quality QA datasets in a short time using crowdsourcing. Crowdworkers can interpret driving blog articles based on the traffic commonsense shared by the society. Secondly, if computers can understand driving situations correctly by extracting driving behavior from blogs, it is possible to predict danger and warn drivers to achieve safer transportation.

Our contributions are summarized as follows.

- We propose an MC-based PAS analysis model and show its superiority to a state-of-the-art neural model.
- We construct PAS-QA and RC-QA datasets in the driving domain using crowdsourcing.
- We improve Japanese PAS analysis by combining the PAS-QA and RC-QA datasets.

[*] The current affiliation is Yahoo Japan Corporation.

Proceedings of the Second Workshop on Machine Reading for Question Answering, pages 98–104
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

2 Related Work

2.1 QA Dataset Construction

FitzGerald et al. (2018) and Michael et al. (2018) constructed QA-SRL Bank 2.0 and QAMRs using crowdsourcing, respectively. They asked crowdworkers to generate question-answer pairs that represent a PAS. These datasets are similar to our PAS-QA dataset, but different in that we focus on omitted arguments and automatically generate questions (see Section 3.1).

Many RC-QA datasets have been constructed in recent years. For example, Rajpurkar et al. (2016) constructed SQuAD 1.1, which contains 100K crowdsourced questions and answer spans in a Wikipedia article. Rajpurkar et al. (2018) updated SQuAD 1.1 to 2.0 by adding unanswerable questions. Some RC-QA datasets have been built in a specific domain (Welbl et al., 2017; Suster and Daelemans, 2018; Pampari et al., 2018).

2.2 Machine Comprehension Models

Many MC models based on neural networks have been proposed to solve RC-QA datasets. For example, Devlin et al. (2019) proposed an MC model using a language representation model, BERT, which achieved a high-ranked accuracy on the SQuAD 1.1 leaderboard as of September 30, 2019.

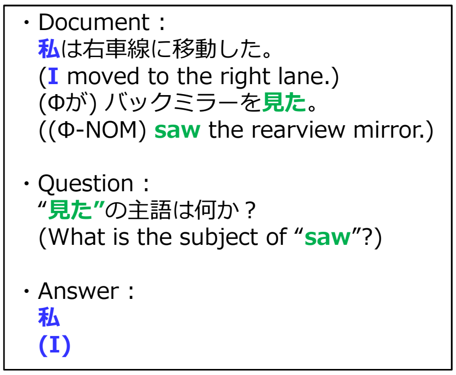
As a previous study of transfer learning of MC models to other tasks, Pan et al. (2018) pre-trained an MC model using an RC-QA dataset and transfered the pre-trained knowledge to sequence-to-sequence models. They used SQuAD 1.1 as the RC-QA dataset and experimented on translation and summarization. While they used different models for pre-training and fine-tuning, we use the same MC model by constructing PAS-QA and RC-QA datasets in the same QA form.

3 QA Dataset Construction

We construct PAS-QA and RC-QA datasets in the driving domain. Both the QA datasets consist of triplets of a document, a question and its answer as in existing RC-QA datasets. We employ crowdsourcing to create large-scale datasets in a short time. Figure 1 and Figure 2 show examples of our PAS-QA and RC-QA datasets.

3.1 PAS-QA Dataset

We construct a PAS-QA dataset in which a question asks an omitted argument for a predicate. We focus on the *ga* case (nominative), the *wo* case (accusative), and the *ni* case (dative), which are targeted in the Japanese PAS analysis literature (Shibata et al., 2016; Shibata and Kurohashi, 2018; Kurita et al., 2018; Ouchi et al., 2017).

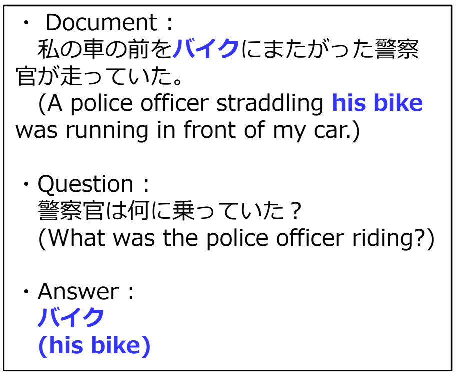


・Document：
私は右車線に移動した。
(**I** moved to the right lane.)
(Φが) バックミラーを**見た**。
((Φ-NOM) **saw** the rearview mirror.)

・Question：
“**見た”**の主語は何か？
(What is the subject of “**saw**”?)

・Answer：
私
(I)

Figure 1: An example of PAS-QA dataset.



・ Document：
私の車の前を**バイク**にまたがった警察官が走っていた。
(A police officer straddling **his bike** was running in front of my car.)

・Question：
警察官は何に乗っていた？
(What was the police officer riding?)

・Answer：
バイク
(his bike)

Figure 2: An example of RC-QA dataset.

As a source corpus, we use blog articles included in the Driving Experience Corpus (Iwai et al., 2019). We first detect a predicate that has an omitted argument of either of the target three cases by applying the existing PAS analyzer KNP[1] to the corpus. KNP tends to overgenerate such predicates, but most erroneous ones are filtered out by the following crowdsourcing step. We extract the sentence that contains the predicate and preceding three sentences as a document. Then, we automatically generate a question using the following template for nominative.

- ［述語］の主語は何か？(What is the subject of [predicate]?)

All the question templates of PAS-QA datasets are shown in Table 1. We ask crowdworkers to choose one from answer choices, which consist of nouns extracted from the document and special symbols,

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

Case	Question
Nominative	［述語］の主語は何か？ (What is the subject of [predicate]?)
Accusative	○○を［述語］、の○○に入るものは何か？ (What is the accusative of [predicate]?)
Dative	○○に［述語］、の○○に入るものは何か？ (What is the dative of [predicate]?)

Table 1: Question templates of PAS-QA datasets.

"author," "other," and "not sure." The details of this procedure are described in the appendix.

We generated questions from 2,146 blog articles. We asked five crowdworkers per question using Yahoo! crowdsourcing[2]. We adopted triplets with three or more votes if they are not "not sure." For accusative and dative PAS-QA questions, we adopted triplets if they are "other." In this case, there is not any antecedent of a zero pronoun in a document, and the answer is "NULL." For nominative PAS-QA questions, we did not adopt triplets if they are "other" because a nominative always exists as a noun in a document or "author." In addition, because "author" is not explicitly expressed in the document, we add a sentence "著者は以下の文章を書きました。" (The author wrote the following document.) to the beginning of the document to deal with "author" in MC models. We record the answers as spans in a document or NULL.

We randomly extracted 100 questions for each case from the PAS-QA dataset and judged whether we can answer them. As a result, 97% nominative, 87% accusative and 68% dative questions were answerable. For accusative and dative, we checked all the questions and chose answerable ones. Finally, we created 12,468 nominative, 3,151 accusative and 1,069 dative triplets including 476 accusative and 126 dative questions whose answers are NULL. It took approximately 32 hours and approximately 210,000 JPY to create this dataset.

3.2 RC-QA Dataset

We construct a driving-domain RC-QA dataset in the same way as SQuAD 1.1. We extract a document from the Driving Experience Corpus and ask three crowdworkers to write questions and their answers about the document. After that, we ask another five crowdworkers to answer a question to validate its answerability and adopt questions with three or more same answers.

	Nominative	Accusative	Dative	Other
# Questions	41	28	8	123
Ratio	20.5%	14.0%	4.0%	61.5%
Ratio (Omission)	(5.0%)	(2.5%)	(0.5%)	−

Table 2: Classification of questions in the RC-QA dataset.

Training method		Dataset
MC-single		PAS-QA
Joint training	MC-merged	PAS-QA + RC-QA
	MC-stepwise	RC-QA → PAS-QA

Table 3: Three training methods for PAS analysis.

As a result, we obtained 20,007 RC-QA triplets from 5,146 blog articles. It took approximately 60 hours and approximately 180,000 JPY to create this dataset.

We randomly extracted 200 questions from the RC-QA dataset and judged the question types. The result is shown in Table 2. A question was classified according to whether it is a question asking for any argument of nominative, accusative or dative, and if applicable, whether it is an omission or not. As shown in Table 2, the RC-QA dataset contains nearly 40% of questions asking arguments of nominative, accusative and dative, and a few questions asking for omitted arguments, which are similar to the PAS-QA dataset. There are various other questions asking for arguments other than nominative, accusative and dative, and questions using why and how.

4 PAS Analysis Based on a Machine Comprehension Model

We analyze PAS based on the MC model on our constructed PAS-QA dataset. Each question in the PAS-QA dataset asks an omitted argument and has an answer that is expressed as a span in the given document or NULL. Because the PAS-QA dataset has the same structure as existing MC datasets including NULL, such as SQuAD 2.0, we can employ an existing state-of-the-art MC model that answers a span in the document or NULL.

We refer to the method of MC training based only on the PAS-QA dataset as **MC-single**. We also propose two joint training methods that use both the PAS-QA and RC-QA datasets: **MC-merged** and **MC-stepwise**, as described in Table 3. The purpose of these joint training methods is to verify whether domain knowledge can be learned from the RC-QA dataset and whether it is

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

	Train	Development	Test
Nominative	11,359	544	565
Accusative	2,756	199	196
Dative	967	50	52

Table 4: Split of the PAS-QA dataset.

effective in improving the accuracy of PAS analysis. In MC-merged, the PAS-QA and RC-QA datasets are just merged and used for training. In MC-stepwise, the RC-QA dataset is used for pre-training, and this pre-trained model is fine-tuned using the PAS-QA dataset.

5 Experiments

We conduct PAS analysis experiments of our MC-single/merged/stepwise methods using the PAS-QA and RC-QA datasets. We also compare our methods with the neural network-based PAS analysis model (Shibata and Kurohashi, 2018) (hereafter, NN-PAS), which achieved the state-of-the-art accuracy on Japanese PAS analysis.

5.1 Experimental Settings

We adopt BERT (Devlin et al., 2019) as an MC model. We split the triplets in the PAS-QA dataset as shown in Table 4. All sentences in these datasets are preprocessed using the Japanese morphological analyzer, JUMAN++[3].

We trained a Japanese pre-trained BERT model using Japanese Wikipedia, which consists of approximately 18 million sentences. The input sentences were segmented into words by JUMAN++, and words were broken into subwords by applying BPE (Sennrich et al., 2016). The parameters of BERT are the same as English BERT$_{\text{BASE}}$. The number of epochs for the pre-training was 30.

The state-of-the-art baseline PAS analyzer, NN-PAS, was trained using the existing PAS dataset, KWDLC[4] (Kyoto University Web Document Leads Corpus), as described in Shibata and Kurohashi (2018). We also trained an NN-PAS model using the PAS-QA dataset in addition to KWDLC (hereafter, NN-PAS$'$). For this training, the PAS-QA dataset was converted to the same format as KWDLC, where questions are deleted, and only answers are used.

The PAS-QA test data is used to compare the baseline methods with the proposed methods. As an evaluation measure, EM (Exact Match) is used for all the MC models. EM is defined as (the number of questions in which the system answer matches the gold answer in the dataset) / (the number of questions in the entire dataset). For each experimental condition, training and testing were conducted five times, and the average scores were calculated.

Training method	PAS	RC	NOM	ACC	DAT
NN-PAS	-	-	0.39	0.38	0.29
NN-PAS$'$	✓	-	0.74	0.45	0.32
MC-single	✓	-	**0.76**	0.52	0.37
MC-merged	✓	✓	**0.76**	0.52	0.43
MC-stepwise	✓	✓	**0.76**	**0.53**	**0.51**

Table 5: PAS-QA test results of MC models and NN-PAS models. “PAS” and “RC” denote the use of the PAS-QA and RC-QA datasets, respectively. “NOM”, “ACC” and “DAT” denote the EM scores of nominative, accusative and dative, respectively.

5.2 Results and Discussion

Table 5 lists evalution results of the NN-PAS models and the MC-single/merged/stepwise models. First, NN-PAS$'$ significantly outperformed NN-PAS, and thus the construction of the domain-specific PAS-QA dataset was effective in domain adaptation of the NN-PAS model. Furthermore, our proposed MC-* models outperfomed NN-PAS$'$. For the joint training models, MC-stepwise was better than MC-single for the accusative and dative cases. MC-merged was inferior to MC-stepwise.

We compared the results of MC-single and MC-stepwise. In examples shown in Figures 3 and 4, only the outputs of MC-stepwise were correct. We found some cases that MC-stepwise successfully captured knowledge in the driving domain. In the example shown in Figure 4, the correspondence between “坂を 上がる” (climb up the slope) and “坂を 越える” (going up the slope) can be recognized. MC-merged’s answer “坂道” (the hill road), which has a coreference relation with “ 坂” (the slope), looked correct although “坂” (the slope) was the only answer from crowdsourcing. Supplying multiple answers considering coreference relations is our future work. From these results, we think that it is important to use an RC-QA dataset to acquire domain knowledge, and suggest that it is better to construct both PAS-QA and RC-QA datasets to develop a PAS analyzer for a new

[3] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN++

[4] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KWDLC

・Document：
彼、「車を返す時ガソリンを満タンにするのか、だいぶ走ったから少なくなったなぁ」
(He said, “When we return this car, we have to fill up with gasoline. We ran a lot so we ran out of it.”)
同乗者、「勿体無いから坂道はニュートラルで走ろうぜ」
(The passenger said, “Because It is a waste to consume gasoline, let's run downhill with the gear in neutral.”)
皆それは良いと、賛同して長い坂をニュートラルで下り始めました、直線の長い坂道でした。
(Everyone agreed that it was good idea, and started to go down a long downhill, which was straight.)
・・・と彼は下りにガソリンは要らないと、エンジンを切り鍵を抜いて皆に『**見せました**』。
(He said that we did not need gasoline to go down, turned off the engine, unlocked the key and “**showed**” it to everyone.)

・Question：
○○を『**見せました**』、の○○に入るものは何か？
(What is the accusative of “**showed**”?)

・Answer：

Correct answer	**：**	**鍵 (the key)**
MC-single	：	坂道 (downhill)
MC-merged	：	車 (this car)
MC-stepwise	**：**	**鍵 (the key)**

Figure 3: An example that is correctly answered by MC-stepwise.

domain.

6 Conclusion

We constructed driving-domain PAS-QA and RC-QA datasets using crowdsourcing[5]. We also proposed an MC-based PAS analysis method. In particular, the stepwise training method based on BERT was the most effective, which outperformed the previous state-of-the-art NN-PAS model. In the future, we will pre-train an MC model based on datasets other than the RC-QA dataset to acquire domain knowledge.

・ Document：
屈伸をしながら気合いを入れ直し坂道に挑む。
(I motivate myself again while bending and stretching, and challenge the hill road.)
坂を越えたらバイク屋がある。
(There is a motorbike shop when going up the slope.)
少し『**上っただけで**』さっきまで引いていた汗が今まで以上に噴き出す。
(Just “**climbing up**” a bit, sweat that stopped until a while ago gushes out more than before.)

・Question：
○○を『**上っただけで**』、の○○に入るものは何か？
(What is the accusative of “**climb up**”?)

・Answer：

Correct answer	**：**	**坂 (the slope)**
MC-single	：	汗 (sweat)
MC-merged	：	坂道 (the hill road)
MC-stepwise	**：**	**坂 (the slope)**

Figure 4: An example that is correctly answered by MC-stepwise.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT2019*, pages 4171–4186.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *ACL2018*, pages 2051–2060.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *ACL2018*, pages 2061–2071.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *ACL2007*, pages 132–139.

Ritsuko Iwai, Daisuke Kawahara, Takatsune Kumada, and Sadao Kurohashi. 2018. Annotating a driving experience corpus with behavior and subjectivity. In *PACLIC 32*, pages 222–231.

Ritsuko Iwai, Takatsune Kumada, Norio Takahashi, Daisuke Kawahara, and Sadao Kurohashi. 2019. Development of driving-related dictionary that includes psychological expressions. In *Proceedings of the 25th Annual Meeting of Natural Language Processing (in Japanese)*, pages 1201–1204.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *LREC2002*, pages 2008–2013.

Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. 2018. Neural adversarial training for semi-supervised Japanese predicate-argument structure analysis. In *ACL2018*, pages 474–484.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *NAACL2018*, pages 560–568.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP2018*, pages 2381–2391.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2017. Neural modeling of multi-predicate interactions for Japanese predicate argument structure analysis. In *ACL2017*, pages 1591–1600.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. A span selection model for semantic role labeling. In *EMNLP2018*, pages 1630–1642.

[5] These datasets are available at http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?Driving%20domain%20QA%20datasets

Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *EMNLP2018*, pages 2357–2368.

Boyuan Pan, Yazheng Yang, Hao Li, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. 2018. MacNet: Transferring knowledge from machine comprehension to Sequence-to-Sequence models. In *NeurIPS2018*, pages 6095–6105.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *ACL2018*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *EMNLP2016*, pages 2383–2392.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL2016*, pages 1715–1725.

Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. 2016. Neural network-based model for Japanese predicate argument structure analysis. In *ACL2016*, pages 1235–1244.

Tomohide Shibata and Sadao Kurohashi. 2018. Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In *ACL2018*, pages 579–589.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *EMNLP2018*, pages 5027–5038.

Simon Suster and Walter Daelemans. 2018. CliCR: a dataset of clinical case reports for machine reading comprehension. In *NAACL2018*, pages 1551–1563.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *AAAI2018*, pages 4929–4936.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *ACL2017*, pages 191–200.

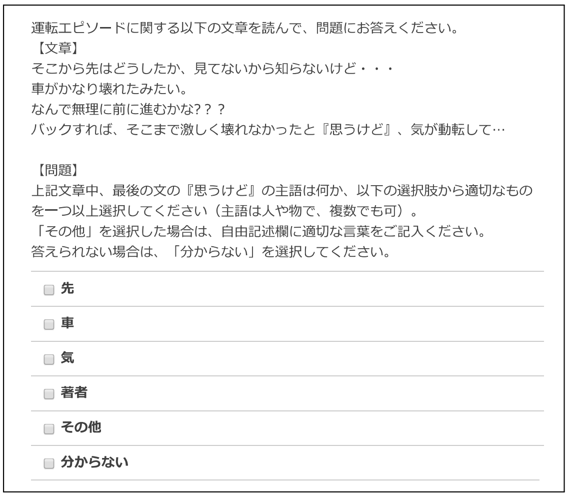
Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *EMNLP2017*, pages 94–106.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP2018*, pages 2369–2380.

A Details of PAS-QA Dataset Construction

We construct the PAS-QA dataset asking for omitted nominative arguments using the following procedure:

1. We extract four consecutive sentences that satisfy the following conditions from the Driving Experience Corpus constructed by Iwai et al. (2019).
   - The Driving Experience extracting CRF tool (Iwai et al., 2018) judges that three or more sentences out of four sentences are driving experience.
   - Each sentence contains at least one PAS.
   - The PAS analyzer, KNP, judges that there is a PAS whose nominative argument is omitted in the fourth sentence.
   - Sentences include at least one "Driving Characteristic Word" (Iwai et al., 2019).
2. We automatically make crowdsourcing tasks using an extracted document and a PAS whose nominative argument is omitted (See Figure 5 and Figure 6). Each task consists of a document, a question and answer choices. Answer choices consist of nouns extracted from the document and special symbols, "author," "other," and "not sure." For nominative PAS-QA questions, the special symbol "author" can often be an answer, but it is not explicitly expressed in the document. So we add it to the choices. We add "other" so that it can be selected when there is an appropriate answer besides the choices. We add "not sure" so that workers can select it if they cannot find an answer. We add more explanations to crowdsourcing answer screen (See Figure 5 and Figure 6).
3. Using crowdsourcing, we ask five crowdworkers per question to select one or more appropriate answers from the choices. We asked five crowdworkers per question using Yahoo! crowdsourcing. We adopted triplets with three or more votes if they are not "not sure." If they are "other," we handled them as described in the main paper. We finally record the answers as spans in a document or NULL.



運転エピソードに関する以下の文章を読んで、問題にお答えください。
【文章】
そこから先はどうしたか、見てないから知らないけど・・・
車がかなり壊れたみたい。
なんで無理に前に進むかな?？？
バックすれば、そこまで激しく壊れなかったと『思うけど』、気が動転して…

【問題】
上記文章中、最後の文の『思うけど』の主語は何か、以下の選択肢から適切なものを一つ以上選択してください（主語は人や物で、複数でも可）。
「その他」を選択した場合は、自由記述欄に適切な言葉をご記入ください。
答えられない場合は、「分からない」を選択してください。

□ **先**

□ **車**

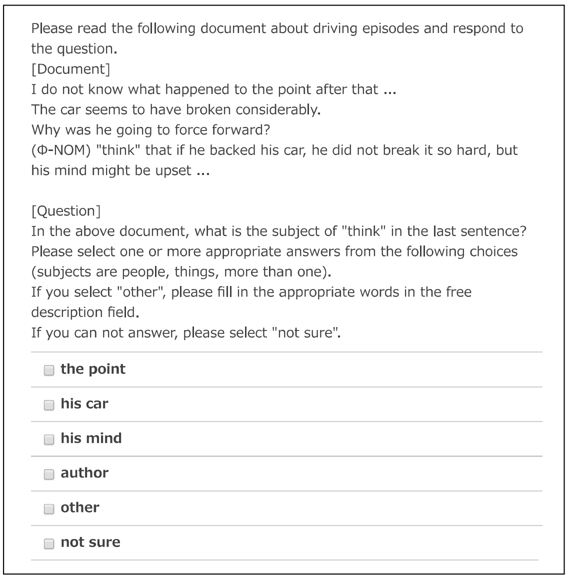
□ **気**

□ **著者**

□ **その他**

□ **分からない**

Figure 5: PAS-QA dataset answer screen.



Please read the following document about driving episodes and respond to the question.
[Document]
I do not know what happened to the point after that ...
The car seems to have broken considerably.
Why was he going to force forward?
(Φ-NOM) "think" that if he backed his car, he did not break it so hard, but his mind might be upset ...

[Question]
In the above document, what is the subject of "think" in the last sentence?
Please select one or more appropriate answers from the following choices (subjects are people, things, more than one).
If you select "other", please fill in the appropriate words in the free description field.
If you can not answer, please select "not sure".

□ **the point**

□ **his car**

□ **his mind**

□ **author**

□ **other**

□ **not sure**

Figure 6: PAS-QA dataset answer screen (English translation version).