Influence of Classifiers and Encoders on Argument Classification in Japanese Assembly Minutes

Yasutomo Kimura¹, Hideyuki Shibuki², Hokuto Ototake³, Yuzu Uchida⁴, Keiichi Takamaru

⁵, Kotaro Sakamoto⁶, Madoka Ishioroshi², Teruko Mitamura⁷, Noriko Kando²,⁸

¹ Otaru University of Commerce, Japan ² National Institute of Informatics, Japan

³ Fukuoka University, Japan ⁴ Hokkai-Gakuen University, Japan

⁵ Utsunomiya Kyowa University, Japan ⁶ Yokohama National University, Japan

⁷ Carnegie Mellon University, USA ⁸ SOKENDAI, Japan

kimura@res.otaru-uc.ac.jp

Abstract

We performed a comparative study of the influence of seven different types of classifiers and four types of encoders on argument classification in Japanese assembly minutes using 45 sets of results from the Question Answering Lab for Political Information task at the NTCIR-14 workshop. The more accurate value obtained from a classification of argumentative relations between a speech sentence and a political topic was 0.942 using the support vector machines classifier and one-hot encoding, while the most accurate classification value obtained with the long short-term memory classifier and word embedding was estimated to be 0.934.

1 Introduction

Numerous arguments about various topics are conducted at different assemblies globally. Although the arguments are valuable for the general public, they are too numerous and intertwined to be comprehensible. Meanwhile, the demand for promptly providing the information required by the users after checking facts to eliminate the fake news from such arguments has been increasing in recent times. Advanced question answering (QA) technologies including argument mining and/or machine comprehension can assist the users to avail such information. Therefore, argument mining from assembly minutes is becoming increasingly significant.

In general, machine learning methods, such as the support vector machines (SVM), are used to recognize argumentative relations, including support or attack relations [Stab and Gurevych, 2017]. However, determining the most suitable method and relevant design of argument vectors for argument mining from assembly minutes is a challenge. The QA Lab-PoliInfo (Question Answering Lab for Political Information) task¹ [Kimura *et al.*, 2019] at the NTCIR-14 workshop was held from January 2018–June 2019. This task was a shared task that focused on recognizing and summarizing the opinions of assemblymen and their reasons in the Japanese Regional Assembly Minutes Corpus [Kimura *et al.*, 2016]. Fifteen teams participated and submitted a total of 119 results. These teams had employed different types of methods, such as the rule-based classifier vs. machine learning classifier, one hot encoding vs word embedding, and SVM vs. long short-term memory (LSTM).

The QA Lab-PoliInfo task includes the segmentation, summarization and the classification tasks. The objective of the classification task is to recognize the classes of the speech of assemblymen, such as "support", "against" and "other", to an opinion, such as "The Tsukiji Market should move to Toyosu area". This is similar to recognizing the argumentative relations. We investigated the influence of the difference between classifiers and encoders in recognizing argumentative relations using the results of the classification task.

The main contribution of this study is to clarify the influence of classifiers and encoders of the machine learning methods on argument classification in the Japanese assembly minutes based on the results of various empirical systems.

2 Related work

The comparative study on argument mining is presented in this section. Aker et al. [Aker *et al.*, 2017] comparatively analyzed the machine learning methods and feature sets using persuasive essays and Wikipedia articles in English. However, the results do not include the current methods, such as the LSTM. The Japanese assembly minutes include different characters from the essays and articles.

Fake News Challenge² and CLEF-2018 Fact Checking Lab³ [Nakov *et al.*, 2018] are shared tasks that deal with political information. The Fake News Challenge conducted the stance detection task and estimated the relative perspective (or stance) of two pieces of text relative to a topic, claim, or issue. The CLEF-2018 Fact Checking Lab conducted two tasks, which consists of the check-worthiness and the factuality [Atanasova *et al.*, 2018; Barrón-Cedeño *et al.*, 2018]. As Japanese arguments are generally more implicit than English, there is some uncertainty about the effectiveness of the argument mining methods for English with respect to Japanese texts.

¹https://poliinfo.github.io/

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²http://www.fakenewschallenge.org/

³http://alt.qcri.org/clef2018-factcheck/



Figure 1: Example of the classification task

Stanford Question Answering Dataset (SQuAD) [Rajpurkar *et al.*, 2016] is used for advanced QA purpose, including machine comprehension [Wang *et al.*, 2018; Wang *et al.*, 2017]. While the SQuAD includes 100,000+ questions, the data set used in the QA Lab-PoliInfo task comprises 10,000+ questions. The latter, therefore, is not capable of providing sufficient amount of training data for general machine learning methods. However, consistently securing sufficient amount of training data is considered difficult in a specific domain like assembly minutes. Researching on the results obtained from limited amount of data is important on account of their execution in the real world.

3 Analysis settings

3.1 Data

The Japanese Regional Assembly Minutes Corpus [Kimura *et al.*, 2016] had collected the minutes of plenary assemblies in 47 prefectures of Japan from April 2011–March 2015. These Japanese minutes resemble a transcript. In the question-and-answer session, an assemblyman asks several questions at a time, and a prefectural governor or a superintendent answers the questions under his/her charge. Any speech is too extensive to understand its contents at a glance; therefore, information access technologies, such as the advanced QA and automated summarization, aid in this process. A subset of the corpus, which was narrowed down to the Tokyo Metropolitan Assembly, was used for the QA Lab-PoliInfo task.

For the gold standard data, 14 political topics, such as "The Tsukiji Market should move to Toyosu area," were considered in advance. After all the sentences including keywords in a topic, such as "Tsukiji Market," were extracted from the corpus, at least three workers annotated the gold standard data per sentence using cloud services. Finally, a total of 10,291 sentences were used as the training data, and 3,412 sentences were used as the test data.

```
{
    "ID":"Classification-2018-JA-FormalSample-00001",
    "ID":"Classification-2018-JA-FormalSample-00001",

 "Topic": "Tsukiji Market should movet to Toyosu area.
 "Utterance":"Toyosu area, I have a high expectation of the
ability to attract customers because new facility
¥"Senkyaku-Banrai¥" will be built. ",
 "Relevance":1
 "Fact-checkability":1,
  "Stance":1,
  "Class":1
}.
 "ID":"Classification-2018-JA-FormalSample-00003".
  "Topic":"Tsukiji Market should movet to Toyosu area. "
 "Utterance":" I am against the relocation of New Bank
Tokyo and Tsukiji market. ",
  "Relevance":1,
 "Fact-checkability":0,
 "Stance":2,
 "Class":0
},
```

Figure 2: Example of data in JSON format

3.2 Classification task

The objective of the classification task at the QA Lab-PoliInfo task is to discover the opinion, which possesses the factcheckable reasons, in the Japanese assembly minutes. Figure 1 shows an example of the classification task. Firstly, a political topic was provided. When a speech sentence in the minutes was provided, the basic factors of classification, which were relevance, fact-checkability and stance agreeing, were recognized. Relevance implies checking whether the sentence provided refers to the specific topic. Fact-checkability implies checking whether the sentence provided contains fact-checkable reasons. Stance agreeing implies checking whether the speaker of the sentence agrees with the topic. However, we prepared a third stance, called "other", to denote that a speaker stands neutral or demonstrates no relation to the topic. Finally, the sentence was classified into the following three classes: support with fact-checkable reasons (S), against with fact-checkable reasons (A), and other (O). All the data are provided to the participants in JavaScript Object Notation (JSON) format, as shown in Figure 2.

As measured from the evaluation, the accuracy of all classes A is defined as follows.

$$A = \frac{1}{|Q|} \sum_{q \in Q} \frac{num(q)}{3} \tag{1}$$

where Q is a set of sentences provided, and num(q) is the number of workers, who annotated the classified class as the gold standard class in the sentence q (maximum value = 3).

Input: A political topic and a sentence in the minutes

Output: A relevance (existence or absence), a factcheckability (existence or absence), a stance agreeing (agree, disagree, or other) and a class (support with factcheckable reasons, against with fact-checkable reasons, or other)

Evaluation: accuracy of all classes

3.3 Grouping the methods

During the classification task, the results of 45 methods from 11 teams were submitted. As the methods were varied, de-

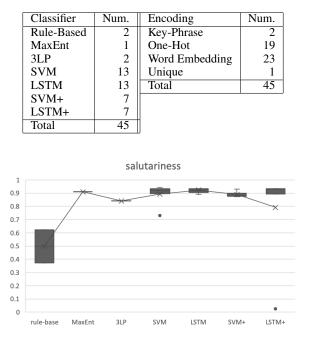


Table 1: Numbers in classifiers and encoders



Table 2: The most accurate classification values obtained from classifiers and encoders

classifier	max	encoding	max
rule-based	0.624	key-phrase	0.624
MaxEnt	0.909	one-hot	0.942
3LP	0.842	word embedding	0.934
SVM	0.942	unique	0.909
LSTM	0.934	-	
SVM+	0.932		
LSTM+	0.933		

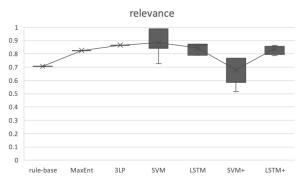


Figure 4: Accuracy of relevance in classifiers

vising an approach to group them was difficult. As the teams also submitted their system description, we grouped the methods according to viewpoints that are shared by many methods, i.e., based on the type of machine learning classifier and encoding.

Although most methods used a machine learning classifier, there were two rule-based methods. Some methods employed a combination of classifiers, such as SVM and decision tree. Therefore, we decided the classifier groups as follows: rule-based, MaxEnt, three-layered perceptron (3LP), SVM, LSTM, a combination of SVM and other classifiers (SVM+), and a combination of LSTM and other classifiers (LSTM+). There was no method that used a combination of SVM and LSTM.

The encoding of the methods using the machine learning classifier was performed through either one-hot encoding or word embedding. However, one method was observed to be an exception, as its encoding included folding a word and its appearing place into a vector element. The rule-based classifiers used simple key-phrases without encoding. Therefore, the encoding groups were decided as follows: key-phrase, one-hot encoding, word embedding, and unique encoding. Table 1 lists the numbers in the classifier and the respective encoding groups.

4 Result

Figures 3–10 show the box-and-whisker plots with respect to the accuracy of classification, relevance, fact-checkability and stance agreeing in the classifier and encoding groups, respectively. Table 2 lists the most accurate of all the values. The accuracy results of the machine learning classifiers were observed to be better than that of the rule-based classifiers. The SVM classifier demonstrated the most accurate value of 0.942, while the LSTM classifier demonstrated a value of 0.934. The combinations of classifiers did not work as well as they were expected. An accuracy of 0.942 with the one-hot encoding was the best, although it was marginally higher than that of word embedding (0.934). Aker et al. [Aker *et al.*, 2017] reported that the difference between the classifiers was marginal, and the results observed in this study exhibited a similar tendency.

While comparing the basic factors of classification with each other, it was observed that the results of fact-checkability were relatively low. As it is an important factor for a wellgrounded argument, it can emerge into an issue in the future.

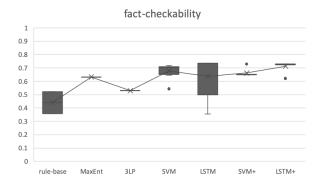


Figure 5: Accuracy of fact-checkability in classifiers

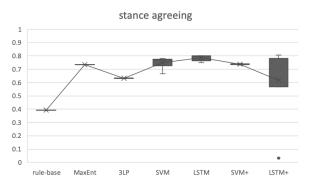


Figure 6: Accuracy of stance agreeing in classifiers

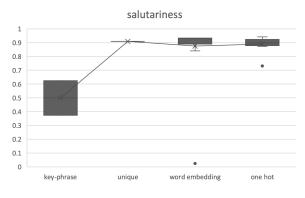


Figure 7: Accuracy of classification in encoders

5 Conclusion

We performed a comparative study of the influence of seven types of classifiers and four types of encoders on argument classification in Japanese assembly minutes using 45 sets of results from the QA Lab-PoliInfo task at the NTCIR-14 workshop. During the classification of argumentative relations between a speech sentence and a political topic, the most accurate value obtained using an SVM classifier and one-hot encoding was estimated to be 0.942. However, the accuracy of the combination of an LSTM classifier and word embedding was estimated to be 0.934.

References

- [Aker et al., 2017] Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. What works and what does not: Classifier and feature analysis for argument mining. In Proceedings of the 4th Workshop on Argument Mining, pages 91–96, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [Atanasova et al., 2018] Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, task 1: Check-worthiness.

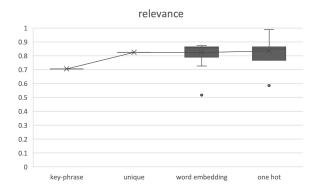


Figure 8: Accuracy of relevance in encoders

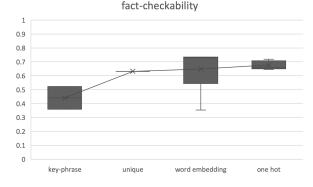


Figure 9: Accuracy of fact-checkability in encoders

In Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors, *CLEF 2018 Working Notes*. *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France, September 2018. CEUR-WS.org.

- [Barrón-Cedeño et al., 2018] Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, task 2: Factuality. In Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors, CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Avignon, France, September 2018. CEUR-WS.org.
- [Kimura et al., 2016] Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu Uchida, Hokuto Ototake, and Shigeru Masuyama. Creating japanese political corpus from local assembly minutes of 47 prefectures. In Proceedings of the 12th Workshop on Asian Language Resources (ALR12), pages 78–85, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.
- [Kimura et al., 2019] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro

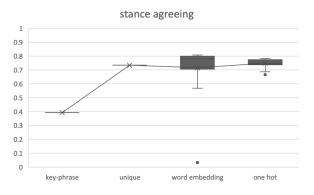


Figure 10: Accuracy of stance agreeing in encoders

Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, Harumichi Yuasa, Satoshi Sekine, and Kentaro Inui. Overview of the ntcir-14 qa lab-poliinfo task. In *Proceedings of the 14th NTCIR Conference*, Tokyo, Japan, June 2019.

- [Nakov et al., 2018] Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In Josiane Mothe Fionn Murtagh Jian Yun Nie Laure Soulier Eric Sanjuan Linda Cappellato Nicola Ferro Patrice Bellot, Chiraz Trabelsi, editor, Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, Avignon, France, September 2018. Springer.
- [Rajpurkar et al., 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.
- [Stab and Gurevych, 2017] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, September 2017.
- [Wang et al., 2017] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 189–198, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [Wang et al., 2018] Wei Wang, Ming Yan, and Chen Wu. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa-

pers), pages 1705–1714, Melbourne, Australia, July 2018. Association for Computational Linguistics.