NAMI Question Answering System at QA Lab-PoliInfo

Ken-ichi Yokote[1] and Makoto Iwayama[1]

Hitachi Central Research Laboratory, Tokyo, Japan
{kenichi.yokote.fb,makoto.iwayama.nw}@hitachi.com
https://www.hitachi.com

Abstract. Argument detection is considered to be a key factor in much previous work related to QALab-PoliInfo Segmentation task. However, It has different views about "argument" thus classifying sentences in terms of argument detection may be a noisy process. In this paper, we propose a method that has filter-by-confidence step after assuming all text segments to be an argument , instead of argument detection step. Our method achieved 93.9 precision and 81.3 recall,indicating that filter-by-confidence is helpful to avoid negative affect of noisy text classification process.

Keywords: argument mining · discourse analysis · text summarization
Team Name
nami
Subtasks
QA Lab PoliInfo Segmentation Task

1 Introduction

The QA-Lab PoliInfo task[10] is a task that is related to argument mining[9]. Most studies of argument mining formulate problems within a framework of (a) searching for arguments by judging sentence types, and (b) searching these arguments for relations(e.g., for the QA-Lab PoliInfo task, summary and original text relation)[5]. However, different studies have made different views about what sort of sentences can be regarded as arguments and what sort of properties these sentences have. Palau et al. assumed a binary classification, whereby a text either contains an argument or it does not [12]. Lippi et al. hypothesized that arguments can be classified into two categories, claims and evidence, each consisting of multiple subcategories[11]. Feng et al. classified arguments into two types: conclusions and assumptions [7]. Duthie et al. postulated that there are two types of arguments, supporting statements and rebuttals [6], and Peldszus et al. postulated that argument texts can be classified along the three axes of rhetorical attributes, evidence-related attributes, and argument-related attributes [13]. In addition to arguments, several studies have also made different claims about the summary texts [8, 14]. Jones et al. classified summary texts into

two types, informative and indicative[8], while Radev et al. classified them into general and topic-oriented [14]. While there are many claims of this sort, it is difficult to know the true nature of the arguments and summary texts provided by QA-Lab PoliInfo task. Therefore, it is not appropriate to use some assumption to formulate (a) as a text classification problem, because the assumption itself may be erroneous. In this study, we proposed a method whereby (a) candidate sentences are searched only based on discourse structure and attribute information without determining the sentence type, (b) multiple methods are used to find related pairs among the candidate sentences, and (c) finally evaluating which of the search results from (b) is most probable by confidence.

2 Proposed method

2.1 System overview

Figure 1 shows an overview of this method. The details of each step will be described later in Section 2.2,2.3,2.4,2.5.(a) corresponds to Section 2.2,2.3. (b) and (c) correspond to Section 2.4,2.5.

First, a new column called " Utterance Segment ID " is provided in the primary information. Although utterances may be placed on adjacent lines in the minutes corpus, this does not necessarily mean they are separated from each other. The utterance segment ID is a column that indicates the separation of utterances with a value that increments whenever the value in the Speaker column changes. In Figure 1, utterance segment ID 714 is assigned to lines 22241 and 22242, while a new utterance segment ID 715 is assigned to lines 22498 and 22499 because the speaker changes from " 宮崎章 " to " 石原慎太郎 ".

【1】Person-Role Relation Detection In person-role relation detection, a minutes corpus is used to associate roles with people in primary information. When the " AnswerSpeaker "value in secondary information is " 知事 " and the " Speaker "value in primary information is " 石原慎太郎 ", it must be recognized that the 知事 refers to 石原慎太郎. As Fig. 1 shows, detecting the person-role relation assigns the " 役職名 " column to the primary information.

【2】Utterance Segment Detection Discourse information indicating who responded to whom is used to extract an area that has correspondence relationship with secondary information (hereinafter defined as a " utterance segment " Consider searching for questions and answers from 宮崎章 and 石原慎太郎. 石原慎太郎 speaks in various parts of the discussion, and not only with 宮崎章. If we can scan all the comments made by 石原慎太郎 and identify the parts where he is speaking to 宮崎章, then we will be able to narrow down the search candidates. As shown in Fig. 1, the discourse relationship between utterance segments

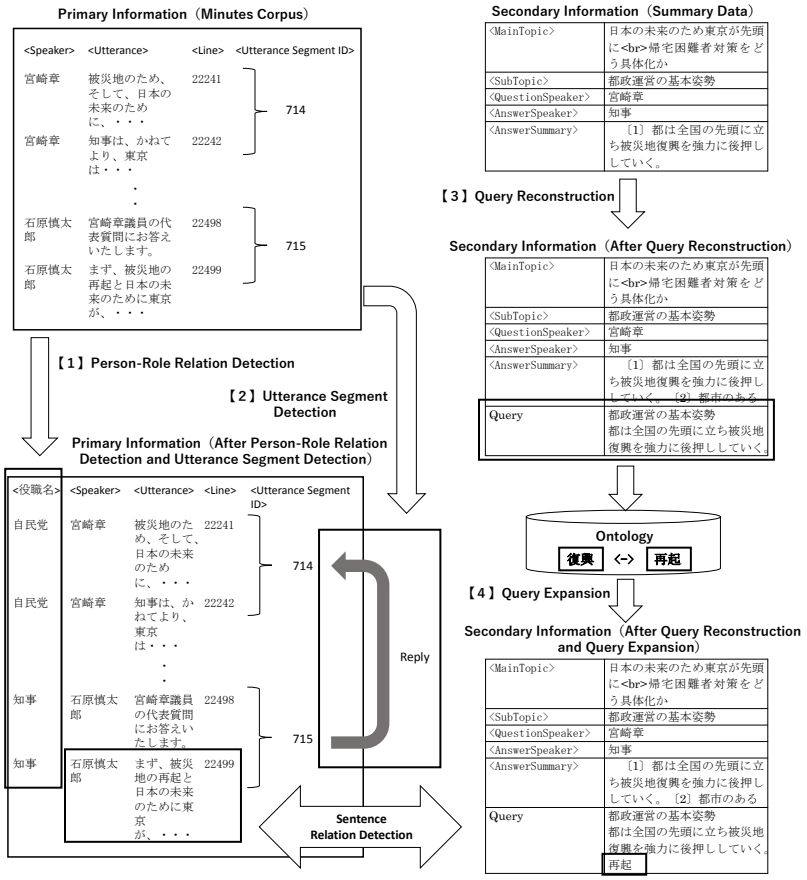

Fig. 1. Example of processing in the proposed method

715 and 714 can be recognized by detecting the utterance segment. Using this information, we can identify that utterance segment 715 has a correspondence relationship with the secondary information.

【3】Query Reconstruction As shown in Fig. 1, the query is generated by appending the AnswerSummary word set to the SubTopic word set by performing Query Reconstruction. Here, appending means combining all the elements of two word sets into a single word set.

【4】Query Expansion In query expansion, a dictionary of related terms (ontology) is used to add synonyms and related terms to the query to cope with variations in the choice of expressions.
Figure 1 shows that Query Expansion added the word " 再起 ", which is related to the word " 復興 " . By adding " 再起 ", we can associate this query with the statement at the start of line 22,499: " まず、被災地の再起と … ".

2.2 Person-Role Relation Detection

In this method, we determined the person-role relations by evaluating the conditional probability of co-occurrence of words in the minutes corpus. Even when referring to the same speaker, the same notation is not necessarily used for the primary and secondary information. Figure 2 shows a concrete example of the problem addressed by this step.

Since the Line of the primary information is 22,754 and the AnswerStarting-Line of the secondary information is 22,754, these two items are in a corresponding relationship. But while the Speaker value is " 前田信弘 ", the AnswerSpeaker value is " 産業労働局長 ", which is not the same. At this step, the name is estimated from the role by the following expression.

$$Speaker(R) = \arg\max_{N_t} P(N_t|R) \tag{1}$$

Here, R represents a job title such as " 産業労働局長 ", and N is a set of person names constructed by aggregating the Speaker column of the minutes corpus. N_t indicates the t-th element of N. Table 1 shows an example of data co-occurrence between a person and a role.

Table 1. Example of co-occurrence between a person and a role

Line	Utterance
26867	〔産業労働局長前田信弘君登壇〕

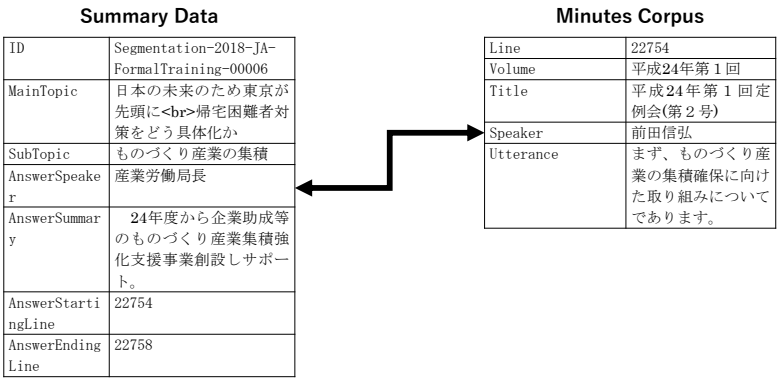

Fig. 2. Example of person-role relation detection

When the keyword " 産業労働局長 " appeared, the keyword " 前田信弘 " also appeared, so it is possible to recognize a correspondence relationship between the two by evaluating the above formula.

2.3 Utterance Segment Detection

In this method, we use heuristics to estimate the discourse relationships between the utterance segments. Table 2 shows the discourse table for " 石原慎太郎 ". " Utterance Segment ID " refers to the utterance segment ID described above in Section 2.1, and " Start Line " and " End Line " indicate the corresponding start line and end line. " QuestionSpeaker " indicates to whose question the speaker in this segment ID is responding. By constructing the discourse table in advance, we can efficiently narrow down the region corresponding to the secondary information. For example, we assume that there is a secondary information in which

Table 2. Example of a discourse table (Speaker: 石原慎太郎)

QuestionSpeaker	Utterance Segment ID	Start Line	End Line
鈴木あきまさ	250	8274	8362
増子博樹	235	7742	7764
大山とも子	283	9297	9352
小磯善彦	268	8881	8922

" AnswerSpeaker "value is " 知事 "and " QuestionSpeaker "value is " 鈴木あき

まさ ”,and we can recognize the relationship between “ 石原慎太郎 ” and “ 知事 ” by using the person-role relationship obtained as described in Section 2.2. If there is no discourse table, then the candidates for “ AnswerStartingLIne ” and “ AnswerEndingLine ” are all the lines corresponding to the segment ID of 石原慎太郎, i.e., 8274 through 8362, 7742 through 7764, 9297 through 9352, and 8881 through 8922. However, by drawing up a discourse table using “ 鈴木あきまさ ”, the candidates can be narrowed down to just one range: 8274 through 8362. The discourse table is constructed by assigning the most recent questioner as seen from each utterance segment ID. This is based on a heuristic that assumes questioners are answered one at a time during the discussions.

2.4 Query Reconstruction

Table 3. Comparison of word break types

Word break type	Dictionary for Mecab	Character normalization
Neologd	Neologd	No
Neologd_NORM	Neologd	Yes
IPA	ipadic	No
IPA_NORM	ipadic	Yes

Word Break In this method, text is divided into words (tokens) using a morphological analysis tool called Mecab [1]. Table 3 shows a classification of word breaking processes. “ Character Normalization ” column refers to the string normalization method of Neologd [2]. In the “ Dictionary for Mecab ” column, “ ipadic ” refers to ipadic dictionary[3]. “ Neologd ” refers to Neologd dictionary[4]. In the following, a word (token) is defined as “ word information ”, and a set of word information is defined as a “ word information list ”.

All-text-segment Search All-text-segment search works as follows. Suppose the “ Line ” columns 10 through 12 are extracted as the utterance segment. In this case, there are six continuous partial areas in line units: (10), (11), (12), (10,11), (10,11,12) and (11,12). These are used as targets for verifying the correspondence with the secondary information. Each element is hereinafter defined as a “ candidate sentence ”.
confidences are used to verify whether each candidate sentence corresponds with the secondary information. confidences are described in the next section

Confidence When a candidate sentence (described in the previous Section) is selected and associated with secondary information, the likelihood of this association is evaluated according to three different indicators. The query chunk size (hereinafter defined as Q_{chu}) is an evaluation indicator that uses only secondary

information. The query coverage (Q_{cov}) and query occupation rate (Q_{occ}) are evaluation indicators that use both primary and secondary information. These indicators are calculated as follows.

$$Q_{cov} = \sum_{i=0}^{|Q|} \frac{Inc(D, Q_i)}{|Q|} \tag{2}$$

$$Q_{occ} = \sum_{i=0}^{|Q|} \frac{Inc(D, Q_i)}{LineNum(D)} \tag{3}$$

$$Q_{chu} = \min |C_t| \tag{4}$$

D and Q are word information lists generated using the wordbreak process described above. D is a word information list generated from the “ Utterance ” column of the candidate sentence, Q is a word information list generated from the secondary information and C is a list of chunks generated from the secondary information. i indicate the index positions of these word information lists. Table 4 shows an example of data with three chunks. Here, Q consists of the questions

Table 4. Example of summary data that has chunks

QuestionSpeaker	QuestionSummary
宮崎章	〔1〕被災地そして日本の未来のため東京は先頭に立つべき。知事の所見は。〔2〕「2020 年の東京」計画に込めた決意は。〔3〕24 年度予算に込めた思いは。

“〔1〕被災地そして日本の未来のため東京は先頭に立つべき。知事の所見は。〔2〕「2020 年の東京」計画に込めた決意は。〔3〕24 年度予算に込めた思いは。”, which are converted into word information lists by regarding them as a single text. On the other hand, C is a word information list that is generated by dividing the questions into three groups: “〔1〕被災地そして日本の未来のため東京は先頭に立つべき。知事の所見は。”, “〔2〕「2020 年の東京」計画に込めた決意は。”, and “〔3〕24 年度予算に込めた思いは。”. For example, C_1 indicates the word information list corresponding to the question “〔1〕被災地そして日本の未来のため東京は先頭に立つべき。知事の所見は。”. $Inc(D, Q_i)$ has a value of 1 when word information Q_i is included in D, and 0 otherwise. $LineNum(D)$ is the lumber of lines in the candidate sentence D.

Sentence Relation Detection using Query Reconstruction and Filter-by-confidence In this method, Q is generated by three methods.

(1) Using the word information list in the “ Summary ” column only
(2) Creating Q by appending the word information list of the“ SubTopic ”column to the end of each chunk in the “ Summary ” column

(3) Creating Q by appending the“ MainTopic ”and“ SubTopic ”word information lists to the end of each chunk in the “ Summary ” column

Next, all three types of Q are used to search for D with the highest confidence. The confidence is evaluated by combining one or more of the query coverage, query occupation rate and query chunk size as described above. Finally, from the maximum confidence of each Q, the set of D and Q having the highest values is defined as $D1$ and $Q1$.

2.5 Query Expansion

In this method, $Q1$ is extended by Ontology. First, word information that is highly related to the word information of $Q1$ is retrieved from the ontology and added to $Q1$. The $Q1$ after this addition is designated as $Q2$. Next, for a certain range centered on $D1$, we search again for the D with the highest confidence again using $Q2$. This result is defined as $D2$, and becomes the final output of the system.

3 Experimental results

3.1 Evaluation of Utterance Segment Detection and Person-Role Relation Detection

We can evaluate Precision and Recall for each $D2$. The extraction of the utterance segment and determination of the person-role relationship are evaluated based on whether or not there exists $D2$ with recall and precision of zero. As described above in Section 2.3, the purpose of determining person-role relations and extracting utterance segments is to identify segment ID. If identification of the segment ID is unsuccessful, the recall and precision will inevitably be zero because the subsequent processing of sections 2.2 and 2.3 will search an area that does not contain a correct answer. Therefore, if the recall and precision are not zero, we can consider the segment ID to have been successfully identified. Table 5 shows the number of $D2$ for each metrics value x. x represents presicion when “ Metrics ”column is P,and recall when “ Metrics ” column is R.
The “ Word Break Type ” column indicates the four types of word breaking processing discussed in section 2.4.

3.2 Evaluation of Query Reconstruction

Table 6 compares the performance of query reconstruction using the confidences described above in section 2.4. In the“ Metrics ”column, F represents F-measure values. The meaning of “ Word Break Type ” is the same as in Table 5. The meaning of Q_{cov}, Q_{occ} and Q_{chu} is the same as in section 2.4, “ $Q_{cov} + Q_{occ}$ ” is the harmonic mean of Qcov and Qocc, and “ $Q_{cov} + Q_{occ} + Q_{chu}$ ” is the weighted sum of “ $Q_{cov} + Q_{occ}$ ” and Q_{chu}.

Table 5. Number of D2s for each metrics value

Word Break Type	Metrics	x=0	0<x<0.2	0.2 ≤x<0.4	0.4≤x<0.6	0.6≤x<0.8	0.8≤x<1	x=1
Nelogd	P	0	0	0	3	4	9	67
	R	0	3	21	27	25	7	0
Nelogd_NORM	P	0	0	0	3	4	9	67
	R	0	3	21	27	25	7	0
IPA	P	0	0	0	2	4	8	69
	R	0	2	22	29	26	4	0
IPA_NORM	P	0	0	0	2	4	8	69
	R	0	2	22	29	26	4	0

Table 6. Query reconstruction performance

Word Break Type	Metrics	Qcov	Qocc	Qcov + Qocc	Qcov + Qocc + Qchu
Nelogd	P	0.5310	0.9508	0.9519	0.9557
	R	0.5917	0.1868	0.5286	0.5400
	F	0.5597	0.3122	0.6798	0.6901
Nelogd_NORM	P	0.5310	0.9508	0.9519	0.9557
	R	0.5917	0.1868	0.5286	0.5400
	F	0.5597	0.3122	0.6798	0.6901
IPA	P	0.5238	0.9536	0.9521	0.9580
	R	0.5928	0.1868	0.5088	0.5190
	F	0.5562	0.3124	0.6632	0.6732
IPA_NORM	P	0.5238	0.9536	0.9521	0.9580
	R	0.5928	0.1868	0.5088	0.5190
	F	0.5562	0.3124	0.6632	0.6732

3.3 Evaluation of Query Expansion

Table 7 compares the performance of query expansion with different confidences. The meanings of the column names are the same as in Table 6. “ No Query Expansion ” indicates the same environment as IPA_NORM in Table 6. “ Query Expansion ” represents the results of appling query expansion to “ No Query Expansion ”. In Section 3.2, since different word breaking types only resulted in a small difference in precision, the query expansions are compared using only IPA_NORM.

Table 7. Query expansion performance

	Metrics	Qcov	Qocc	Qcov + Qocc	Qcov + Qocc + Qchu
No Query Expansion	P	0.5238	0.9536	0.9521	0.9580
	R	0.5928	0.1868	0.5088	0.5190
	F	0.5562	0.3124	0.6632	0.6732
Query Expansion	P	0.5746	0.9157	0.9261	0.9396
	R	0.8478	0.5059	0.8052	0.8137
	F	0.6850	0.6517	0.8614	0.8721

References

1. http://taku910.github.io/mecab/
2. https://github.com/neologd/mecab-ipadic-neologd/wiki/Regexp.ja
3. http://sourceforge.jp/projects/ipadic/
4. https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md
5. Cabrio, E., Villata, S.: Five years of argument mining: a data-driven analysis. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. pp. 5427–5433. International Joint Conferences on Artificial Intelligence Organization (7 2018). https://doi.org/10.24963/ijcai.2018/766, https://doi.org/10.24963/ijcai.2018/766
6. Duthie, R., Budzynska, K., Reed, C.: Mining ethos in political debate. In: COMMA. pp. 299–310 (2016)
7. Feng, V.W., Hirst, G.: Classifying arguments by scheme. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. pp. 987–996 (2011)
8. Jones, K.S.: Automatic summarising: Factors and directions. In: Advances in Automatic Text Summarization. pp. 1–12. MIT Press (1998)
9. Kimura, Y.: Ntcir-14-qalab-poliinfo-2ndroundtablemtg, https://poliinfo.github.io/NTCIR-14-QALab-PoliInfo-2ndRoundTableMTG.pdf
10. Kimura, Y., Shibuki, H., Ototake, H., Uchida, Y., Takamaru, K., Sakamoto, K., Ishioroshi, M., Mitamura, T., Kando, N., Mori, T., Yuasa, H., Sekine, S., Inui, K.: Overview of the ntcir-14 qa lab-poliinfo task. In: Proceedings of the 14th NTCIR Conference (2019)

11. Lippi, M., Torroni, P.: Argument mining from speech: Detecting claims in political debates. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
12. Palau, R.M., Moens, M.F.: Argumentation mining: the detection, classification and structure of arguments in text. In: Proceedings of the 12th international conference on artificial intelligence and law. pp. 98–107. ACM (2009)
13. Peldszus, A., Stede, M.: Joint prediction in mst-style discourse parsing for argumentation mining. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 938–948 (2015)
14. Radev, D.R., Hovy, E., McKeown, K.: Introduction to the special issue on summarization. Computational linguistics **28**(4), 399–408 (2002)