Annotating Online Civic Discussion Threads for Argument Mining

Gaku Morio
Graduate School of Engineering
Tokyo University of Agriculture and Technology
Koganei, Tokyo, Japan
morio@katfuji.lab.tuat.ac.jp

Katsuhide Fujita
Institute of Engineering
Tokyo University of Agriculture and Technology
Koganei, Tokyo, Japan
katfuji@cc.tuat.ac.jp

***Abstract*—Argument mining techniques have become popular in online civic discussion thread analysis to understand an enormous amount of posts and flow of discussions for consensus building. However, the existing corpora and discussion thread analysis haven't discussed argument mining schemes sufficiently. This paper proposes a novel scheme for discussion thread analysis, annotates online civic discussions, and analyzes the annotated corpus. Our scheme consists of novel inner- and inter-post schemes. The inner-post scheme considers a post as a stand-alone discourse in a thread. We perform a micro-level annotation of argument components and relations in a post. The inter-post scheme provides a micro-level inter-post interaction to capture the argumentative reply-to relation. As a result, we have an annotated corpus including 399 threads and 5559 sentences of 204 citizens that is valid and argumentative. In addition, we analyze the annotated corpus to demonstrate statistical and linguistic properties of the corpus.**

***Index Terms*—argument mining, online civic engagement, computational argumentation, annotation, corpus, inter-annotator agreement**

I. Introduction

Online civic discussions support consensus building among hundreds of citizens without time or location limitations [1]–[4]. To analyze such large-scale online civic discussions, computational argumentation (CA) and argument mining (AM) [5], [6] derived from the argumentation theory are employed to understand an enormous amount of posts and flow of the discussions. Being distinct from opinion mining, AM focuses on structuring of argument components (AC), such as premises and claims [7], which enables us to understand proofs and evidences behind opinions. Therefore, CA and AM techniques are expected to be an effective tool in online civic discussion analysis.

CA is a corpus driven discipline that analyzes natural language discourses. In particular, AM is one of the CA disciplines focusing on extracting claims or premises and inferring their structures from a discourse. Stab et al. [8] argue that the task of AM is divided into the following three subtasks:

- **Component identification** focuses on separation of argumentative and non-argumentative text units and identification of AC boundaries.

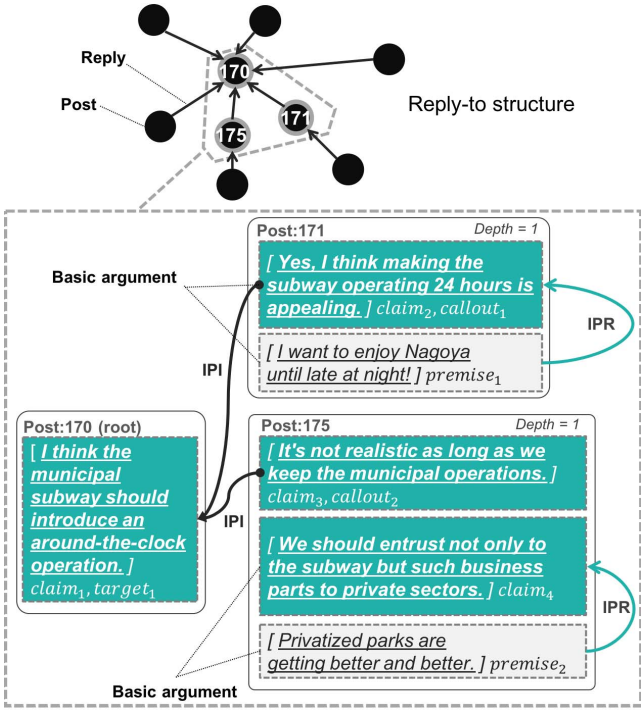


Fig. 1: Example annotation of discussion threads (The actual annotation of discussion threads is done in Japanese). The top graph represents a reply-to structure of the threads. Each post has arguments (possibly empty). An argument consists of a claim and possibly empty set of premises. Inner-post relation (IPR) indicates an inner-post relation between ACs. Inter-post interaction (IPI) indicates an inter-post interaction of two ACs between a post and its replying post.

- **Component classification** addresses the function of ACs. It aims at classifying ACs into different types, such as claims and premises.
- **Structure identification** focuses on linking arguments or ACs. Its objective is to recognize different types of argumentative relations, such as support or attack relations.

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.00-39

The structure identification can also be divided to macro- and micro-level approaches. Generally, identifying structures among ACs is categorized as a micro-level approach, and among complete arguments as a macro-level approach. Stab et al. [8], [9] annotated micro-level structures of ACs. They utilized the three subtasks and proposed "one-claim" approach for annotating persuasive essays. To summarize, they represented an essay as a set of arguments. An argument is a tree composed of a claim as a root node and (possibly empty) set of premises as leaves. In comparison with existing macro-level approaches, the authors focus on the micro-level annotation, therefore, many CA related studies utilize these datasets [10]–[13], [13]–[15].

While the necessity of corpora is argued, not enough annotation schemes for discussion thread analysis in AM have been developed. This is due to the shortage of tools for making a conclusive micro-level annotation of discussion threads. Recently, Ghosh et al. [16] defined a target and callout scheme that features reply-to relations. However, their scheme addresses relations between complete arguments and ignores the micro-structure of arguments between posts. Another example is that of Hidey et al. [17], who annotated the micro-level structure of ACs. However, they considered an entire thread as a discourse. Thus, they allowed a premise that links to a claim in another post, while a post should be considered as a stand-alone discourse because a writer for each post is different. Therefore, we cannot simply utilize their scheme. In addition, the annotated data from [17] with only 78 threads makes it more difficult to employ state-of-the-art discrimination methodologies like [18]–[21].

Motivated by the demands for online civic discussion analysis and weaknesses of both discussion thread schemes and corpora, we have developed a novel corpus of online civic discussion threads for AM. Our main contributions are as follows:

1) To remedy the shortage of the micro-level annotation for discussion threads, a novel micro-level combination scheme is proposed. We introduce *inner-post* and *inter-post* schemes in combination. This combination enables us to discriminate arguments per post, rather than per thread as in [17]. In the former scheme, a post is assumed as a stand-alone discourse and a micro-level annotation is provided. In the second scheme, we introduce inter-post micro-level interactions. The introduction of the micro-level interactions allows us to capture more informative argumentative relations between citizens than [16].
2) To utilize our proposed scheme in practice, annotations to an actual large-scale online civic discussion data with thread structures are achieved. Threads (399) and sentences (5559) are annotated by 11 annotators. To the best of our knowledge, this work is the first attempt to annotate such large-scale civic discussions for AM.[1] In addition, we evaluate inter-annotator agreement scores to assure the reliability of our annotated corpus. Consequently, we have found that our annotation with the proposed scheme is valid.
3) Our annotated corpus is analyzed in detail to reveal statistical and linguistic properties of the corpus. Some useful findings, both common properties in AM and unique properties in our corpus, are shown.

[1]Recently, Park et al. [22] provided a similar dataset of civic engagement, while their dataset doesn't consider post-to-post relations sufficiently.

The remainder of this paper is organized as follows. First, we describe related research and datasets. Next, the novel annotation study for online civic discussion analysis is proposed. Then, we analyze the annotated corpus statistically and linguistically. Finally, we present conclusion and future work.

II. Related Works

Corpus for Computational Argumentation

A dataset is essential for AM since it is a corpus driven discipline. There are some annotated corpora for AM. For instance, epoch-making AraucariaDB [23] contains heterogeneous document types like news articles and online discussions. The authors annotated reasoning types and implicit argument components. However, they did not evaluate the inter-annotator agreement. Thus, the corpus is open for questions in terms of reliability. An annotation that focuses on blog comments also exists [16]. The authors of the later work focus on interactions between posts to remedy the infancy of studies on online interactions in AM. Annotations for 2016 US Presidential Debates [24] and analysis of *ChangeMyView*[2] in Reddit are further examples of discussion corpora however, they mainly concern with extremely argumentative and logically obvious discourses. Hence, these studies differ from our study in which we tackle casual and amateur online civic discussions between hundreds of citizens.

Argument Structure Identification

There are two types of approaches to identify the structure of argumentation in documents [8]. The first one is a macro-level approach as in [16], [25], [26]. This approach addresses relations between complete arguments and ignores the micro-structure of arguments [8]. In [16], the authors introduced a scheme to represent relations between two posts by *target* and *callout*; however, their study discards micro-level structures in arguments because of their macro-level annotation. The second one is a micro-level approach as in [6], [8], [9], which focuses on the relations between ACs. In [6], arguments are considered as trees. In [8], the authors also represented relations of ACs in essays as tree structures. However, they addressed discourses of a single writer (i.e., an essay writer) rather than multiple authors in a discussion thread. Therefore, we can't simply apply their scheme to our study since each participant may have different standpoints in an online discussion. Moreover, this scheme lacks of micro-level approaches for reply-to structure.

The number of citizens: 204, likes: 4326, browse histories: 21993 , threads:399, posts:1327, tokens:120241, the number of posts per thread:3.33 $\pm$ 3.29, the depth of thread:1.09 $\pm$ 1.19, the number of sentences per post:4.19 $\pm$ 3.33, the number of words per sentence:21.63 $\pm$ 19.92.

TABLE I: Statistics of COLLAGREE discussions. $\pm$ indicates a standard deviation.

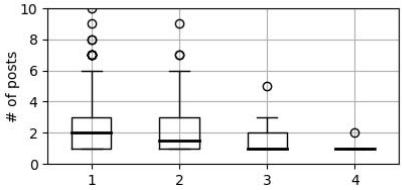


Fig. 2: The number of posts per depth.

III. Data

In this study, we concentrate on annotating online civic discussions based on an actual online civic discussion data from COLLAGREE [1], [3], [4], which is an online civic forum. The discussion theme was "the charm of Nagoya City." The discussion was held from the end of 2016 to the beginning of 2017, and co-hosted by the government of Nagoya City and the Nagoya Institute of Technology. One topic per one thread is assigned, hence, we regard each thread is independent. TABLE I summarizes statistics of our collected data. The accumulated data includes 399 threads, 1327 posts, 5559 sentences and 120241 tokens in total. Thus, we have many threads compared to [17] who annotated only 78 threads. The average number of sentences per post is 4.19, indicating that a post typically contains a lot of information. Fig. 2 in turn illustrates a transition of the number of posts according to the thread depth. It shows that the majority of posts are posted in depth 1 or 2.

IV. Annotation Study

A. Problem Setting

Modeling argumentation can be recognized as the transformation from an informal logic of natural language text to a structured representation as argument diagramming [27]. While there are some studies on transformations, most of them propose to transform from argumentations to trees [28] because the tree structure has a particular affinity for learning or optimizations. Therefore, this particular structure is introduced to our inner-post scheme.

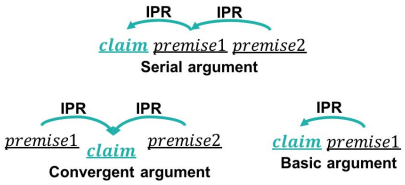
B. The Combination Scheme

A key factor to apply AM to thread structures is combining schemes pursuant to existing studies, since there are many existing solid AM schemes. In TABLE I, the property that a post has a satisfactory number of sentences shows one post usually argues claims or premises in a way. Regarding a post as a stand-alone discourse is thus presumed to be appropriate rather than conducting the annotation of [17] that doesn't regard a post as a stand-alone discourse. In this work, we present a novel scheme combining *inner-post* scheme of a stand-alone post with *inter-post* scheme that considers a reply-to argumentative relation.

1) Inner-Post Scheme: We model the inner-post scheme which is similar to [9]. We consider a post as (possibly empty) a set of arguments. The structure of each argument is modeled with a "one-claim" approach.[3] The approach considers an argument as the pairing of a single claim and a (possibly empty) set of premises that justify the claim. We define ACs (i.e., claim and premise) and IPR as follows:

- **Claim** is a controversial statement and the central component of an argument, and has no outgoing links.
- **Premise** is a reason for justifying (or refuting) the claim.
- **Inner-post relation (IPR)** is a directed argumentative relation in a post describing the relationship that one component has with another. Each IPR:$(target \leftarrow source)$ indicates that the $source$ component is either a justification for or a refutation of the $target$ component. Thus, a $source$ should be a premise, and each premise has a single outgoing link to another premise or claim [19]. Therefore, only two patterns, $(claim \leftarrow premise)$ and $(premise \leftarrow premise)$ exist. Note that the looped relations are forbidden. While [8] annotated support and attack labels, our preliminary annotation result indicates that support relations are in the majority. Thus, this study doesn't consider introducing the support and attack scheme.

In addition, to maintain a tree structure, one argument should have a claim as a root node and to be connected in an acyclic graph. For better understanding, we show three examples of arguments:



A basic argument is composed of a claim and a premise that supports the claim; a convergent argument comprises of two premises that support the claim individually; an argument is serial if it includes a reasoning chain. A structure with one-claim and multiple (more than two) premises are surely derived from the combination of the arguments. This study employs the inner-post scheme with the arguments described above to execute a micro-level annotation.

2) Inter-Post Scheme: Next, we provide an inter-post scheme to capture the relation in reply-to structures. Especially, there is a need to capture relations between claims of

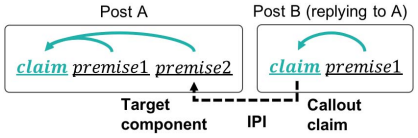
[2]https://reddit.com/r/changemyview

[3]Though [8] introduced major claim as a consistent claim of an essay; we do not introduce the scheme because a post has too little arguments to define the major claim.

citizens when visualizing arguments in online civic discussions [2]. Hence, we incorporate an interaction model [16] into our scheme for inter-post relations. The authors introduced macro-level relations of complete arguments with a target and callout scheme. However, we propose micro-level relations of ACs in the reply-to structure to combine the micro-level inner-post scheme with the interaction model. We define the target, callout, and micro-level IPI as follows:

- **Target** is a prior component that has been called out by a subsequent claim in another post that replies to the post of the target.
- **Callout** is a subsequent component that refers back to the prior target. In addition to referring back to the target, a callout explicitly includes stance. Hence, a callout should be a claim and each callout has one outgoing link to its replying post.[4]
- **Inter-post interaction (IPI)** is the inter-post relation of two posts: a parent post and a child post that replies to the parent post. A relation $(parent \leftarrow child)$ represents that the $child$ is a callout and the $parent$ is a target.

As we define above, a callout is restricted to be a claim. For example, in the following structure, the post B replies to post A, only the claim of post B can be a callout as follows:



Note that a target can be of any component type: claim, premise, or non-argumentative. Consequently, employing the inter-post scheme, as described in this section, captures the micro-level reply-to argumentative relations.

Fig. 1 denotes an example of the proposed inner- and inter-post scheme for a discussion thread. The graph in the top of the figure represents a reply-to structure [26] of the thread. Due to the limitations of space, we focus on only the three posts. For example, post 175 has a basic argument, and the link between $target_1$ of post 170 and the $callout_2$ of post 175 is an IPI.

C. Annotation Process

The annotation task performed by trained annotators includes three subtasks that [29] identify as part of the AM problem: (1) segmentation, (2) segment classification, and (3) relation identification. This paper follows this definition of the subtasks. Despite Stab et al. [8] employed a manual token-level annotation for the segmentation, the manual token-level segmentation is hard because it requires extensive human resources, time, and cost to build a gold standard. Therefore, we apply a rule-based technique for the segmentation. Then, we split a comment in a post into sentences [30] and consider each sentence as a component [31]. That is, one sentence becomes one component. Next, for the segment classification, the component type, AC (claim or premise) or non-AC (non-argumentative (NA)), for each component is annotated. Finally, the relation identification is required to annotate IPRs and IPIs. To conclude, the annotation subtask (1) is conducted automatically, while the subtasks (2-3) are conducted manually.

[4]To restrict a callout to a claim makes our problem more simple because the number of outgoing links from a claim becomes one at a maximum. Thus, we introduced the restriction.

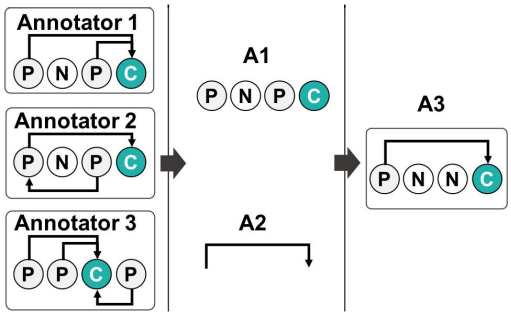


Fig. 3: Compiling a gold standard in the first phase of annotation by three independent annotators, where P, C, and N represent premise, claim and non-argumentative respectively.

Using multiple phases for multiple annotation subtasks is common [8], [9], [32]. To annotate our data, we provide two phases. In the first phase, we focus on annotating component types and IPRs, and create a partial gold standard. At the beginning of the first phase, we ask the annotators to read the entire thread as knowing a topic and stance of a discourse improves inter-annotator agreement [9]. In the second phase, IPIs are annotated based on the partial gold standard. Thus, each annotator annotates IPIs (targets and callouts) based on the same ACs and IPRs of the partial gold standard.

Next, we explain the detailed annotation process. In our study, a majority of the vote is employed to create the gold standard. The three annotators have annotated independently. The detailed procedure of the first phase for compiling the partial gold standard is as follows:

A1: Each component type (premise, claim, or NA) is decided by a majority vote. When the component type of the sentence cannot be decided by majority vote, NA is assigned to them.
A2: Each IPR (link existence) is decided by a majority vote.
A3: Merging the results from A1 and A2, and obtaining trees where root is a claim. Thus, we have trees corresponding to the number of claims in a post.
A4: Eliminating premise tags that do not belong to any tree, assigning them to NA, and eliminating their IPR.

Fig. 3 demonstrates an annotation example performed by three independent annotators. The second phase annotation in the same manner as A2 (replacing IPR to IPI) is as follows:

A5: For the partial gold standard obtained in A4, each IPI (link existence) is decided by a majority vote.

Chunk	Claim	Premise	NA	IPR	IPI
1	54.8	57.2	61.4	37.9	38.7
2	53.9	57.0	50.6	44.6	46.0
3	47.0	52.1	47.0	41.2	42.1
4	56.7	55.3	52.4	44.1	45.0
Avg.	53.1	55.4	52.9	42.0	42.9

TABLE II: Inter-annotator agreement of our annotation.

		size	avg. per post	SD
general	Threads	399	-	-
	Posts	1327	-	-
	Sentences	5559	4.19	3.33
comp	Claims	1449	1.09	0.67
	Premises	2762	2.08	2.36
	NAs	1348	1.02	2.19
rel	IPRs	2762	2.08	2.36
	IPIs	745 ($C \leftarrow C$: 574, $P \leftarrow C$: 109, $NA \leftarrow C$: 62)	0.56	0.62

TABLE III: Statistics of the final corpus. SD means a standard deviation. $(C \leftarrow C)$ denotes that both the target and callout are claims, $(P \leftarrow C)$ denotes that the target is a premise and callout is a claim.

After the subtask A5, the final gold standard is achieved. All 399 threads are annotated using the process from A1 to A4 in the first phase and A5 in the second phase.

D. Inter-Annotator Agreement

To examine a reliability of our annotation, inter-annotator agreement (IAA) scores are evaluated as corpora with an extremely low agreement have no reliability in practice and no persuasion in discussions. In our scheme, IAA scores for component type in the segment classification, and IPR or IPI in the relation identification are evaluated. We employ Fleiss's κ [33] that is popular in AM annotations [7], [8], [17] to evaluate the agreement reliability.

TABLE II shows our IAA scores. Our dataset is divided into four chunks. For each chunk, we assign three trained student annotators. Consequently, 11 annotators engaged in our annotation. As distinct from the essay dataset with relatively formal writings [8], our corpus exhibits a more realistic proportion of badly-structured writings resulting in a relatively low agreement. However, classification tasks can still be applied since [34] refers to the κ value from 41.0 to 61.0 as "moderate agreement", which is similar to the result of Hidey et al. [17]. Also, it is clear that the annotators are hard to agree on the IPI subtask with 42.9 in κ comparing to [16] who achieved $64.0+$ in α_U in their macro-level annotation. However, it is thought to be due to the unique property of our corpus since we use a micro-level annotation.

V. Corpus Analysis

TABLE III shows an overview of the final annotated corpus. It contains 4211 ACs (1449 claims and 2762 premises), 1348 non-AC sentences, 2762 IPRs, and 745 IPIs. We also analyze them with post-level statistics. Such a large proportion of

Fig. 4: Histogram of Pos of premises and claims in posts with more than two sentences.

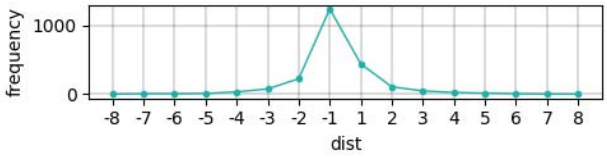


Fig. 5: Histogram of premise $Dist_C$.

premises in comparison with claims is common in argumentative texts since writers tend to provide several reasons for ensuring a robust standpoint [8], [35]. In addition, each post has 1.09 claims and 2.08 premises on average, meaning that treating a post as a stand-alone discourse is a sound assumption. In the case of IPIs, 51% $(745/1449)$ of claims are callouts and 77% $(574/754)$ of targets are claims. It means that our inter-post interactions are pretty argumentative.

To analyze the micro-level inner-post annotation, we examine distributions of claim and premise properties. Herein, Pos is described as a one-dimensional positive position in a post, and $Dist_C$ as a one-dimensional relative position from a claim C. For instance, Pos and $Dist_C$ of the following post with one claim and two premises can be described as:

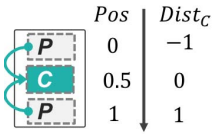
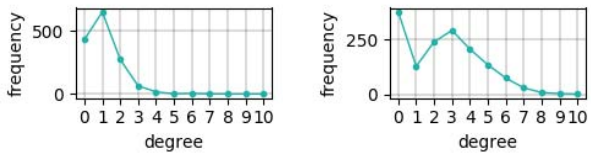


Fig. 4 shows two histograms of Pos of claims and premises with a step of 0.25. Interestingly, claims tend to appear at the end of a post, while premises uniformly appear in any Pos. Thus, the citizens are likely to conclude with their claim after several premises.

Fig. 5 shows a histogram of $Dist_C$. It shows that premises are likely to appear immediately prior to a claim because mode value is -1. In fact, the result exhibits the same property on the essay corpus [19]. Thus, our corpus achieves the reproduction.

Next, we examine claim's *degree* to show how the micro-level claim and premise structure is annotated. The degree means the number of corresponding premises of a claim. For example, the degrees of $claim_1$, $claim_2$ and $claim_3$ in Fig. 1 are 0, 1, and 0 respectively. Fig. 6a shows the resulting histogram. It indicates that the minimum number of arguments with single premise is a mode value. That is, many citizens



(a) COLLAGREE corpus. (b) Essay corpus.

Fig. 6: Histogram of claim's degree for the two corpora. Horizontal axis indicates the claim's degree: 0 means claims without corresponding premises are counted, 1 means claims with one corresponding premise are counted, and 2 means claims with two corresponding premises are counted.

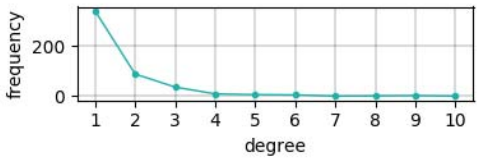


Fig. 7: Histogram of the target's degree. Horizontal axis indicates the target's degree: 1 means targets with one corresponding callout are counted and 2 means – with two corresponding callouts.

state claims without providing any justifications for them. Fig. 6b shows the results from the essay corpus [8] for comparison. The essay corpus shows a different distribution: the mode value is 0 and 3, 2, and 4 subsequently, indicating that a claim tends to have more than 2 premises once at least one premise is linked to the claim. Thus, the property of claim's degrees in online civic discussions differs from the essays.

To study how claims are connected to a target, we show the target's degree in our corpus. In this case, the degree means the number of callouts that are linked to a target. For example, the degree of $target_1$ in Fig. 1 is 2. Fig. 7 illustrates the resulting histogram. This figure shows that the histogram of the target's degree is a power distribution. Many of all targets have less than three callouts. In addition, as shown in Fig. 2, the majority of all targets appear in the root post since deeper posts are less likely to be replied.

Lexical Analysis

This section examines two lexical analyses for our corpus. First, to examine the lexical differences between claims and premises, the occurrence rates of each part-of-speech (POS) of the two ACs are evaluated. Fig. 8 shows the resulting rates of each POS (i.e., verb, noun, adjective, adverb, auxiliary verb, and symbol) over the premise and claim components. As a claim describes author's standpoints, verbs and adjectives usually emerge in claims. Additionally, top three frequent verbs in claims are "し *(do)*, 思い *(think)*, する *(do)*", while "し *(do)*, い *(be)*, ある *(exist)*" in premises, indicating that

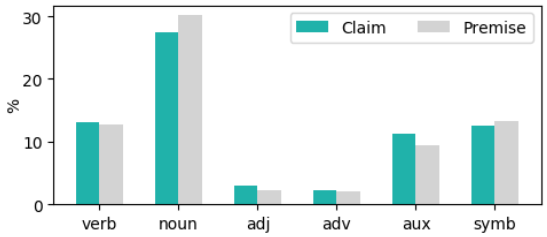


Fig. 8: Proportions of each part-of-speech (POS). For each POS, the significant of averages. at $p < 0.05$ using Mann-Whitney U Test.

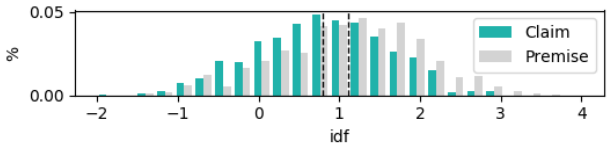


Fig. 9: Histogram of the average IDF value per AC (claim and premise) with more than 5 words. The significance of averages at $p < 0.0001$.

statement verbs can be found in claims more often that in premises. On the contrary, nouns and symbols make an appearance in premises more often than in claims since a premise generally describes knowledge.

Second, we focus on featuring words in claims or premises to show the difference of the use of nouns between the two ACs. In [36], the authors provide an index of expertness by featuring inverse document frequency (IDF). As we presume premises specifically apt to state knowledge compared to claims, premises should have more highly professional nouns. To verify the assumption, we focus on claims and premises with IDF values of nouns. Herein, the IDF of a word w is calculated as follows:

$$idf_w = \log \frac{N}{df\,(w)}, \tag{1}$$

where N represents the number of all components in our corpus and $df\,(w)$ denotes the number of components that contain the word w. Fig. 9 shows the resulting histogram of the average IDF value of nouns per premise or claim component. The figure indicates that the higher IDF of nouns, professional words, appear in premises compared to claims. Specifically, the top four high IDF nouns are "旅客 *(traveler)*, *ID*, 就任 *(accedence)*, 天照大御神 *(Amaterasu-ōmikami: a deity of the Japanese myth cycle)*", while the worst four are "の *(' s)*, 名古屋 *(Nagoya)*, こと *(thing)*, 魅力 *(charm)*." This means that frequently used easy nouns have lower IDF values, and nouns exhibiting expertise have higher IDF values. Therefore, employing IDF values as a feature could be effective for

feature	component classification				link extraction	
	Claim	Premise	NA	Macro-F1	IPR	IPI
BoW	54.2	55.7	58.9	56.3	18.9	10.7
BoW + Pos	54.8	56.0	58.2	56.4	18.7	11.4
BoW + POS	53.5	56.2	56.9	55.5	19.1	10.7
BoW + IDF	54.6	57.0	59.0	56.9	18.5	10.5

TABLE IV: Classification F1 scores (%) on average for each model. The classifier is a linear support vector machine (SVM). BoW: bag of words; POS: part-of-speech; Pos: the positive position in a post; IDF: inverse document frequency in a component. For instance, BoW + POS means we use only bag of words and part-of-speech as features.

discriminating premises and claims, or estimating competence of a citizen.

A. Preliminary Experimental Results for Computational Argumentation

Our ultimate goal is to apply the AM technique to our data, i.e, to automatically classify component types, IPRs and IPIs. To examine the availability of machine learning techniques for the classifications, this section describes a classification experiment. In this work, we provide three features: POS distribution, Pos, and average IDF value. Bag of words (BoW) of top frequent 128 words is also provided as a baseline. We employ support vector machine (SVM) and cross validation. Using SVM for the classifications is important because the classifier is frequently used in many AM studies [8], [15]. The COLLAGREE dataset employed in the experiment is divided into training and testing sets at the ratio of $8:2$. We evaluate F1 scores of classification results for the provided features.

TABLE IV summarizes the performances for each model. Surprisingly, our models perform better than expected in spite of the low κ agreements. The result that simple features achieved 18.9% in the IPR discrimination task is particularly notable, even though the task includes negative-dominated cases. The model that combines BoW with Pos also performs better in the component classification and IPI extraction tasks. Therefore, such structural information is an effective feature for IPI extractions. Alongside of the Pos feature, the IDF feature outperforms the baseline in terms of the component classifications. However, it doesn't outperform in link extractions. The result makes sense because IDF only features lexical information rather than structural information. Also, the part-of-speech distribution feature proves to be useful partially for the link extractions.

VI. Conclusion

This study showed how an AM annotation can be applied to online civic discussion threads. Extending the existing research, we presented a novel combination scheme. In particular, micro-level inner- and inter-post schemes were combined. The proposed inner-post scheme provided a micro-level approach for the argument component – claim and premise – and inner-post relation in a post. The proposed inter-post scheme utilized a reply-to interaction – target and callout – in micro-level. The actual large-scale online civic discussions with 399 threads were annotated. The annotation results showed that these discussions provide a pretty argumentative corpus and an online discussion corpus for AM of a state-of-the-art size. The analysis of our corpus indicated that a claim tends to appear at the end of a post, and premises usually come immediately prior to a claim. A unique property of our corpus is that the number of premises linked to a claim is usually small in comparison with another corpus. In addition to the corpus analysis, we proved that a machine learning technique (SVM in our case) utilizing our proposed features is applicable in component classification and link extraction tasks.

Possible future work includes enhancing our classification model for it to specialize in thread structures, e.g., employing an integer linear programming (ILP) model for end-to-end AM [15]. We will also study applications for future automatic facilitation and consensus building based on AM.

Acknowledgments

This work was supported by CREST, JST (JPMJCR15E1), Japan and JST AIP-PRISM Grant Number JPMJCR18ZL, Japan. We thank Takayuki Ito, Eizo Hideshima, Takanori Ito and Shun Shiramatsu for providing us with the COLLAGREE data.

References

[1] T. Ito, Y. Imi, T. Ito, and E. Hideshima, "Collagree: A faciliator-mediated large-scale consensus support system," in *Proceedings of the 2nd International Conference of Collective Intelligence*, 2014.

[2] T. Ito, "Towards agent-based large-scale decision support system: The effect of facilitator," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[3] G. Morio and K. Fujita, "Predicting argumentative influence probabilities in large-scale online civic engagement," in *Companion Proceedings of The Web Conference 2018*, ser. WWW '18. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018, pp. 1427–1434.

[4] T. Nishida, T. Ito, T. Ito, E. Hideshima, S. Fukamachi, A. Sengoku, and Y. Sugiyama, "Core time mechanism for managing large-scale internet-based discussions on collagree," in *2017 IEEE International Conference on Agents (ICA)*. China: IEEE CPS, July 2017, pp. 46–49.

[5] M. Lippi and P. Torroni, "Argumentation mining: State of the art and emerging trends," *ACM Trans. Internet Technol.*, vol. 16, no. 2, pp. 10:1–10:25, Mar. 2016.

[6] R. M. Palau and M.-F. Moens, "Argumentation mining: The detection, classification and structure of arguments in text," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ser. ICAIL '09. New York, NY, USA: ACM, 2009, pp. 98–107.

[7] J. Eckle-Kohler, R. Kluge, and I. Gurevych, "On the role of discourse markers for discriminating claims and premises in argumentative discourse," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 2236–2242.

[8] C. Stab and I. Gurevych, "Parsing argumentation structures in persuasive essays," *Computational Linguistics*, vol. 43, no. 3, pp. 619–659, 2017.

[9] ——, "Annotating argument components and relations in persuasive essays," in *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, 2014, pp. 1501–1510.

[10] H. V. Nguyen and D. J. Litman, "Contextaware argumentative relation mining," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2016, pp. 1127–1137.

[11] ——, "Argument mining for improving the automated scoring of persuasive essays," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 2018.

[12] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 1882–1891.

[13] D. Ghosh, A. Khanam, Y. Han, and S. Muresan, "Coarse-grained argumentation features for scoring persuasive essays," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 549–554.

[14] I. Persing and V. Ng, "Modeling argument strength in student essays," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 543–552.

[15] ——, "End-to-end argumentation mining in student essays," in *Proceedings of Human Language Technologies: The 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2016, pp. 1384–1394.

[16] D. Ghosh, S. Muresan, N. Wacholder, M. Aakhus, and M. Mitsui, "Analyzing argumentative discourse units in online interactions," in *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA*, 2014, pp. 39–48.

[17] C. Hidey, E. Musi, A. Hwang, S. Muresan, and K. McKeown, "Analyzing the semantic types of claims and premises in an online persuasive forum," in *Proceedings of the 4th Workshop on Argument Mining*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 11–21.

[18] P. Potash, A. Romanov, and A. Rumshisky, "Here's my point: Joint pointer architecture for argument mining," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 1375–1384.

[19] S. Eger, J. Daxenberger, and I. Gurevych, "Neural end-to-end learning for computational argumentation mining," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 11–22.

[20] A. Søgaard and Y. Goldberg, *Deep multi-task learning with low level tasks supervised at lower layers*. Association for Computational Linguistics, 2016, vol. 2, pp. 231–235.

[21] M. Miwa and M. Bansal, "End-to-end relation extraction using lstms on sequences and tree structures," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016, pp. 1105–1116.

[22] J. Park and C. Cardie, "A corpus of erulemaking user comments for measuring evaluability of arguments," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, 2018.

[23] C. Reed and G. Rowe, "Araucaria: Software for argument analysis, diagramming and representation," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 04, pp. 961–979, 2004.

[24] J. Lawrence and C. Reed, "Using complex argumentative interactions to reconstruct the argumentative structure of large-scale debates," in *Proceedings of the 4th Workshop on Argument Mining*. Association for Computational Linguistics, 2017, pp. 108–117.

[25] F. Boltužić and J. Šnajder, "Back up your stance: Recognizing arguments in online discussions," in *Proceedings of the First Workshop on Argumentation Mining*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 49–58.

[26] A. Murakami and R. Raymond, "Support or oppose?: Classifying positions in online debates from reply activities and opinion expressions," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 869–875.

[27] T. Govier, *A Practical Study of Argument*. Belmont: CA: Wadsworth, 1985.

[28] A. Peldszus and M. Stede, "Joint prediction in mst-style discourse parsing for argumentation mining," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 938–948.

[29] ——, "From argument diagrams to argumentation mining in texts: A survey," *Int. J. Cogn. Inform. Nat. Intell.*, vol. 7, no. 1, pp. 1–31, Jan. 2013.

[30] R. Kitagawa and K. Fujita, "Automatic summarization considering time series and thread structure in electronic bulletin board system for discussion," in *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, July 2016, pp. 681–686.

[31] C. Kirschner, J. Eckle-Kohler, and I. Gurevych, "Linking the thoughts: Analysis of argumentation structures in scientific publications," in *Proceedings of the 2nd Workshop on Argumentation Mining*. Denver, CO: Association for Computational Linguistics, June 2015, pp. 1–11.

[32] R. A. Meyers and D. Brashers, "Extending the conversational argument coding scheme: Argument categories, units, and coding procedures," *Communication Methods and Measures*, vol. 4, no. 1-2, pp. 27–45, 2010.

[33] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[34] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, 1977.

[35] R. Mochales and M.-F. Moens, "Argumentation mining," *Artif. Intell. Law*, vol. 19, no. 1, pp. 1–22, Mar. 2011.

[36] G. Morio and K. Fujita, "Competence estimation: Classifying expertise of web discussion participants," in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, July 2017, pp. 801–807.