Competence Estimation: Classifying Expertise of Web Discussion Participants

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

Katsuhide Fujita
Institute of Engineering
Tokyo University of Agriculture and Technology
Koganei, Tokyo, Japan

***Abstract*—Web discussion bulletin boards have been attracting much attention in recent years. They enable people to hold discussions online, whereas these have traditionally been conducted face-to-face at town meetings, etc. However, in a large-scale bulletin board, not all of the participants have a deep understanding of a topic when engaging in discussion. In particular, it is important to automatically classify a person who is making useful posts if a bulletin board is intended to make agreements. This paper proposes an automated method to identify the expertise of participants by defining "expertise" as a requisite argumentative competence. In this paper, we propose novel features for competence estimation models: lexical features (IDF, discourse marker, and topic similarity) and a directed influence graph feature. Furthermore, in the evaluation experiments, we evaluate the precision-recall curve against the baseline. As for datasets for evaluation, the expertise of the participants in the data of discussion conducted in the actual Web discussion bulletin board is annotated in seven grades. The experimental results demonstrate that the proposed methodology is effective in many cases.**

I. Introduction

The Web discussion bulletin board may be effectively applied to town meetings on the Web [1]. Such town meetings mainly cover topics that require local government-level policymaking, and general citizens hold discussions and make agreements on these topics. However, in the case of face-to-face town meetings, the age and status of participants could be uneven or it could be difficult to have a reasonable number of participants due to limitations of time and location. Furthermore, due to a lack of participants, there is a concern that such meetings cannot be used as a tool to make agreements. As a result, Web-based town meetings such as "COLLAGREE" [2], [3], where anybody can participate regardless of time and location, have been emerging.

As there is no limitation of time and location for such a discussion bulletin board and anybody can make a post, it is expected to enable large-scale discussions among a large number of participants. However, with more participants, the number of irrelevant and useless posts increases. In particular, it is important to automatically identify a person who is making useful posts if a Web discussion bulletin board is intended to make an agreement. For example, as shown in Fig. 1, a set of discussion bulletin board participants contains persons with high expertise on the topic, those who have the ability to lead discussion, and some with both traits. These participants tend to influence whether or not agreement can be formed in the discussion. Above all, participants' learning levels of the topic are important elements to form appropriate agreements.

Fig. 1: Existence of set of participants with high competence in set of discussion participants

This paper proposes a method to conduct an automatic classification, by defining an expertise as argument competence. Then, we propose a competence estimation model that sets discourse markers and argumentative structures utilizing the argument mining method [4] as a feature. As far as we know, a competence estimation method utilizing argumentative structures in a discussion bulletin board does not exist. Therefore, this study will contribute to the field of argument mining. Also, as datasets for evaluation, the expertise of participants in several discussions conducted on COLLAGREE is annotated in seven grades. In the evaluation experiments, classification accuracy was evaluated by making the baseline and the proposed method learn by Support Vector Machine (SVM). Also in the evaluation experiments, we make the SVM considering existing and proposed features learn by cross-validation and evaluate them by the precision-recall curve. As datasets for evaluation, the expertise of the participants in the data of discussion conducted in the actual Web discussion bulletin board is annotated in seven grades. Experimental results demonstrate that the accuracy of the proposed method was better than that of the baseline in specific cases, and the proposed method is effective for automatic classification of expertise of Web discussion bulletin board participants.

This paper is organized as follows. First, related studies and the definition of argument competence will be explained. Second, use of the lexical feature and directed influence graph will be proposed. Then, the method of the annotation process and the results of evaluation experiments will be shown. Finally, we will conclude the paper and touch on future work.

II. Related Works

Argument Mining

Attention is being focused on argument mining as an integrated field of data mining, machine learning and argument analysis [4]. Argument mining is a relatively new effort to analyze argumentative structures automatically based on corpuses [5].

Classification of Participants by Golder et al.

By targeting the participants of the Usenet Newsgroup, Golder et al. classified the participants of online communities into the following six categories: Celebrity, Newbie, Flamer, Lurker, Troll, and Ranter [7]. Celebrity is the most important type with regard to agreement formation. Among the participants, Celebrity is expected to have the largest impact on others, write many comments, play a role in leading discussion, and influence other issues, too. In addition, Celebrity surpasses others in expertise in comments and social skills.

Discourse Marker

Eckle-Kohler et al. conducted a study on the role of discourse markers in argumentative discourse [9]. The authors evaluated the classifiable characteristics of the discourse marker against the premise and claim of a discussion by using various statistical methods. As a result, it became evident that the discourse marker is useful for differentiation of premise and claim. As a specific method, with the purpose of evaluating it by modelization, the authors considered that argumentative discourse is an array of rhetorical claims to persuade readers, and then defined that argumentative discourse consists of claim and premise in accordance with the existing studies and experience. Also, they defined that a premise can be classified as one that either supports or attacks the claim. The authors call this minimum unit model the claim-premise model (discussion model or discussion unit model).

Graph Structuring of Argument

Barkar et al. proposed an issue-centered scheme for argument analysis and graph representation of online news. They showed how a summary that grasps argumentative characteristics of the readers' comments is created from a graph [6]. When creating a graph showing a reply relationship in the proposal, Assertions, Viewpoints, and Issues were defined. The comments and reply message of each user are defined as a node, while relationships between rationale to assertion and assertion to issue are defined as the edges.

Influence Diffusion Model

Matsumura et al. focused on text communication and proposed the influence diffusion model [14]. To be specific, modeling was conducted on a process in which a word written by a person will spread to other users by way of the reply relationship. In addition, when a lexical feature is dominant in a community, the original sender of the topic is defined as an opinion leader.

III. Argument Competence Estimation Method Based on Lexical Feature and Argumentative Structures

A. Argument Competence

In this paper, a person with high competence is defined as a participant with high commenting expertise and social skills, in accordance with the definition on competence of online discussion participants by Golder et al. Expertise refers to a competence that is necessary to hold a professional discussion on an issue. Social skills refer to communication capability that is required to promote smooth discussion. Among the two competences, this paper will focus on expertise. This is because the ultimate purpose of a town meeting on the Web, etc. is decision-making or agreement formation. In other words, whether or not a discussion can ultimately lead to an agreement is considered to depend on discussion quality and participants' competence. Finally, expertise is defined as follows.

Argument Competence (Expertise) – A person with a high degree of expertise in the discussion refers to one who has rich knowledge of the discussion theme; and a person with a lower degree has less knowledge.

B. Flow of Automatic Evaluation Method

In this section, the flow of automatic evaluation method will be explained. There are mainly two types of feature extraction: **lexical feature** and **directed influence graph feature**. As to lexical features, feature quantity will be obtained from Inverse Document Frequency(IDF), discourse marker, and topic similarity. As to the directed influence graph feature, feature quantity will be obtained from centrality focusing on reply relationship and influence diffusion model. Finally, a classifier will be created by making the Support Vector Machine (SVM) learn these feature quantities through the use of cross-validation.

In the actual system, when a natural language sentence is given as an input, it will be converted into a string of word class or Bag of Words by using the Japanese morpheme analyzer because words in Japanese documents are not separated by spaces. The data created will become an input data of feature extractors, and it will become a real vector that will be obtained by converting it into a lexical feature or graphic feature. By using it as input data, the SVM classifier is conducted with cross-validation.

C. Lexical Features

The lexical feature refers to the lexical property of messages posted in the Web discussion bulletin board. In this paper, we propose the following three features.

IDF: When estimating the expertise of the Web discussion bulletin board participants, the difficulty level of words used by participants is taken into account. This paper assumes that rareness of words indicates the difficulty level of the words. By using the Inverse Document Frequency (IDF), it will evaluate the words that appear in a small number of sentences rather than the words that appear in a large number of sentences. IDF without smoothing is shown in the following formula:

TABLE I: Word class assuming discourse marker

Word Class	
1	Conjunction
2	Supplementary particle
2	Linking particle
3	Connective particle
4	Parallel marker
5	Adnominal particle
6	Conjunction with postpositional particle, etc.

$$idf_t = \log \frac{N}{df(t)}. \tag{1}$$

N of Formula (1) means the total number of documents, and $df(t)$ means the number of documents including a word t. Furthermore, to the participants $\mathbf{u}_k = (u_1, u_2, .., u_n)$ of discussion topic k, where n persons participate, real vector $\mathbf{c}_k = (c_1, c_2, .., c_n)$, which includes each participant's feature as an element that will be defined. For example, if taking $\mathbf{c}_k$ as IDF value of words, feature quantity c_i of participant u_i will be shown as follows by using the word set W_i, which is Bag of Words of u_i.

$$c_i = \sum_{w \in W_i} idf_w. \tag{2}$$

In this paper, Formula (2) is proposed as a feature of the Web discussion bulletin board participants.

Argument Unit: Eckle-Kohler et al. indicated that a discourse marker in argumentative discourse is effective for separating premise and claim [9]. A premise is an element that supports or attacks a claim, and its structure combined with a claim is called as argument unit. In the study, it becames evident that the words expressing substitute, reason, and continuance, such as "or, as, and," appear significantly often in the premise. On the other hand, words expressing an emphasis, compromise, and comparison, such as "quite, therefore, though, as," tend to appear often in the claim. This paper assumes that classification by a discourse marker is possible for Japanese documents in the same manner and that the frequency of appearance of premise words will indicate the amount of knowledge, which is expertise. Therefore, this study focuses on the word classes shown in Table I, as these are likely to be related to the discourse markers of premise. As to word classes in Table I, this paper uses the normalized number of appearances as the feature quantity.

Topic Similarity: This feature will identify whether or not the comments by the Web discussion bulletin board participants are in line with the discussion topic. The reason for using the similarity to the discussion topic as a feature is that a person with high expertise in a certain topic is not necessarily familiar with another topic. Since it is easy to subclassify expertise, it is considered that a person with high expertise can be also subclassified easily. For example, a person who is familiar with geographical information may not know about a political problem in the same manner. Therefore, this paper will adopt the similarity of a discussion bulletin board's topics and participants' comments as a proposed method.

In recent years, the similarity evaluation of documents has been conducted based on the topic model assuming the latent topic. There are various types of topic models, and the Latent Dirichlet Allocation (LDA, [10]) and the Correlated Topic Models (CTM, [11]) are the leading topic models. LDA is a method that adopted the probability model assuming that topic distribution will follow Dirichlet distribution. It enabled extraction of latent topic inherent in words based on the unsupervised learning. This paper also tried to assume various discussion topics and calculate the similarity against the vectorized data by using LDA. However, in the preliminary experiment, it turned out that there was no correlation at all between the correct data and the estimated data. It was caused by the lack of data. Stated another way, as LDA is a method that generally uses the Markov chain Monte Carlo methods, its learning will largely depend on the number of documents and the number of sample words. In addition, the topic model has a problem in that its effectiveness of perplexity, a general evaluation method, is questioned [12]. Therefore, as an estimation method of the similarity between documents without using a probability model, we have decided to use cosine similarities of TF-IDF. As a method to estimate the similarity by using TF-IDF, Rinott et al. used it as a context-dependent element for the purpose of ranking evidence [13]. TF-IDF is a multiplication of Term Frequency (TF) and IDF, and TF means for the frequency of word appearance by document. IDF is represented by Formula (1) explained above. If the number of appearances of the word t_i in a sentence d_j is represented as $n_{i,j}$, TF-IDF is shown as below:

$$tfidf_{i,j} = tf_{i,j} idf_{t_i} = \frac{n_{i,j}}{\sum_k n_{k,j}} idf_{t_i}. \tag{3}$$

Therefore, similarities between documents can be calculated by using cosine similarity since the vector of TF-IDF value can be created for each document. This paper assumes that the discussion topic is given. As feature quantity, it calculates the cosine similarity between the discussion topic and the posted documents of each user. Table II is an example of topic documents and a document determined by the owner of the Web discussion bulletin board.

D. Directed Influence Graph Feature

The argument competence of participants can be estimated by focusing on the contents of the participants' comments, especially those in the reply sentences. Therefore, this paper proposes the graph and centrality as well as feature quantity based on the influence diffusion model by Matsumura et al. [14]. In the study by Barkar et al. [6], a graph of argumentative structures based on an issue-centered scheme was made manually. However, it is not practical as it costs too much. Thus, this paper will focus on the reply relationship, reply documents and the sender's documents. By doing so, the feature will be automatically extracted against the Web discussion bulletin board with reply structure.

Eigenvector Centrality based on IDF Influence: As previously seen, this feature assumes latent influence of expertise. The reply structure is shown in the directed graph, and feature quantity is obtained by calculating the centrality in the graph. First, we will formulate the influence diffusion centrality in

TABLE II: Examples of Discussion Topic and Topic Sentences

Discussion Topic	Topic Sentences
Environment in Nagoya	A city where people can live pleasantly in a comfortable urban environment. A pleasant city where people can enjoy a rich natural environment. A comfortable city where urban and natural environments coexists with harmony.
Disasters in Nagoya	A city that is resistant to disasters such as earthquakes and heavy rain. A city that is resistant to the occurrence of crimes and accidents. A city where communities are making integrated efforts to secure safety.
Human rights in Nagoya	A city where human rights and bonds are cherished. A city where the elderly and the disabled can live independently without worry. A city where everyone can live energetically in a way he/she likes.
Charms of Nagoya	A city that makes you wish to live, full of charms and vigor. A city that makes you wish to visit, full of charms and vigor. A city its citizens can be proud of.

the reply structure. The directed graph $G := (V, E)$ (user graph) is introduced; this sets the Web discussion bulletin board participants as nodes V and the reply relationship as edges E. As one of the simplest feature quantities in the user graph, it considers the index that counts the number of replies made. In other words, it considers that the more the person received replies, the more he/she has some sort of features. As a centrality based on such a way of thinking, there is a degree centrality [15]). This is a model that takes the number of edges adjacent to a node as the centrality index. However, the degree centrality has a major flaw in that it does not consider the centrality of an adjacent node. For example, as to two user nodes with one connecting edge, let us assume that one is a reply from a user with low degree centrality while the other is a reply from a user with high centrality. In such cases, the degree centrality will evaluate as if both have the same values. In light of the characteristics of discussion, it is necessary to consider the centrality of adjacent user nodes in the discussion bulletin board.

As a method to complement the defect of degree centrality, there is an eigenvector centrality [16]. Eigenvector centrality is a centrality that takes a degree and connecting node into consideration. Eigenvector centrality is calculated by obtaining the eigenvector of the transposed matrix in the adjacent line concerning a user graph's node. In other words, if assuming $A(i, j)$ as below:

$$A(i,j) = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases}$$

Then, eigenvector centrality $c_e(v)$ of user node v will be shown as below:

$$c_e(v) = \lambda^{-1} \sum_{t \in V} A(t,v)\, c_e(t). \tag{4}$$

λ is a constant number. However, as it is a weighted directed graph, the weight of the edge from node i to node j is defined as $w(i, j)$. Furthermore, for $w(i, j)$, weight obtained by expanding the influence diffusion model will be used. From reference [14], in the structure where user node y sends back comments to user node x, comments of x are set as the word set W_x expressed as Bag of Words, and influence $I_{x,y}$ propagated to the user node y, who is making a reply to it, is set as follows:

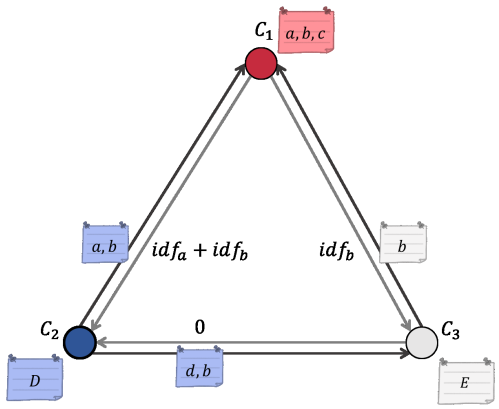

Fig. 2: Example of user graph using IDF influences as weight

$$I_{x,y} = \frac{|W_x \cap W_y|}{|W_y|}. \tag{5}$$

However, since Formula (5) does not take the weight of words into account, words with a higher rate of appearance become noise and less likely to reflect the difference in participants' profiles [17]. Therefore, by giving low evaluation to the words of frequent appearance, IDF, which excels in characterizing words, will be used as the weight. This paper defines it as IDF influence. In other words, as to Formula (4), eigenvector centrality c_{evidf} using IDF influence as weight is shown as follows:

$$c_{evidf}(v) = \lambda^{-1} \sum_{t \in V} A(t,v)\, w(t,v)\, c_{evidf}(t), \tag{6}$$

$$= \lambda^{-1} \sum_{t \in V} \frac{c_{evidf}(t) \sum_{w \in W_t \cap W_v} idf_w}{\sum_{w \in W_t} idf_w}. \tag{7}$$

Fig. 2 is an example of a user graph using IDF influences, where each user has one comment at a maximum and the corpus has only five words – "a, b, c, d, e." For example, comments of user node C_1 consist of three words – "a, b, c" and C_2 and C_3 make respective replies in Bag of Words (BoW) such as "a, b" and "b." The influence C_1 gives to C_2 is shown as the intersection of BoW of C_1, "a, b, c" and BoW of reply comment, "a, b." More specifically, it is the total of IDF value of two words, "a, b" and shown as $idf_a + idf_b$. Similarly,

TABLE III: κ Interpretation of agreement

κ	Interpretation
< 0	Poor agreement
$0.01 - 0.20$	Slight agreement
$0.21 - 0.40$	Fair agreement
$0.41 - 0.60$	Moderate agreement
$0.61 - 0.80$	Substantial agreement
$0.81 - 1.00$	Almost perfect agreement

the influence C_1 will give to C_3 is shown as idf_b. As time series is not taken into account this time, if more than one reply structures from v to t exist, they will be handled as one reply structure. In this paper, Formula (7) is defined as the eigenvector centrality based on IDF influence, and set as the feature quantity.

IV. Evaluation Experiments

A. Annotation Methodology

In our annotation process, five annotators evaluated participants' expertise with regard to the target data based on the definitions in this paper. As to evaluation scales, we carried out a seven-grade questionnaire evaluation with the highest granularity in the Likert scale [8]. In the questionnaire evaluation, we set the standard under which 1 means the lowest competence while 7 means the highest competence. The target data for annotation was the datasets (Number of participants: 187, Number of posts: $1,447$) of six discussion topics taken up in the Web discussion bulletin board, COLLAGREE.

Statistical Information of Annotated Data: When carrying out annotation, five students from the Tokyo University of Agriculture and Technology conducted seven-grade evaluations of the above competence definition (expertise) on datasets of six discussion topics. For each participant, we provided the prior explanations on the points, purpose, and evaluation index of annotation. Then, the agreement of annotation results was evaluated. As a method to evaluate the agreement of annotation results by three or more examiners, Fleiss' Kappa [18] was used, as well as Landis et al.'s proposed index as in Table III on interpretation of the agreement κ [19].

We calculated Fleiss' Kappa for the annotated data and the result was $\kappa = 0.18$. Since this falls under "Slight agreement" based on Table III, it is not likely to be an effective classification task. Therefore, we decided to classify persons with high expertise grades, "6 and 7," as well as low expertise grades, "1 and 2." In other words, we decided to consider two patterns of classification tasks as follows.

Classification Task $\mathcal{H}$
: Among medians of annotation data, classification task of persons with "high" expertise where "7 and 6" are set as *Positive* and "5, 4, 3, 2, and 1" are set as *Negative*.

Classification Task $\mathcal{L}$
: Among medians of annotation data, classification task of persons with "low" expertise where "1 and 2" are set as *Positive* and "3, 4, 5, 6, and 7" are set as *Negative*.

The reason for classifying the patterns as above is to effectively extract persons with high competence and low competence in the Web discussion bulletin board. In particular, a facilitator is required to pay attention to participants' competence level, since a person with high expertise can play a leading role, while a person with low expertise is likely to make useless or inappropriate comments. However, when calculating the medians for the annotated data, very few persons were evaluated with the highest grade, "7," or the lowest grade, "1." Therefore, we relaxed the conditions. As a matter of fact, when the conditions were extended as in the classification task $\mathcal{H}$ and classification task $\mathcal{L}$, the number of labels assigned were as shown in Table IV. In addition, the agreement of the tasks became "Fair agreement" and over as in Table V. Therefore, this paper will analyze these two classification patterns.

TABLE IV: Number of labels in each classification

Classification task	*Positive*/*Negative*	Number of labels (persons)
$\mathcal{H}$	*Positive*	31
	Negative	156
$\mathcal{L}$	*Positive*	44
	Negative	143

TABLE V: Fleiss' Kappa agreement of each classification task

classification task	κ
$\mathcal{H}$	0.33
$\mathcal{L}$	0.51

The results in Table IV show that persons with high expertise are relatively minor in the Web discussion bulletin board. Furthermore, there are a larger number of persons with low expertise compared with persons with high expertise. Therefore, it can be said that only a small number of participants are making contributions in terms of expertise.

B. Results of Evaluation Experiments

By using the feature proposed in this paper, we will make the targets learn the annotated data and evaluate the precision and recall [20]. The following two will be used as the baseline features.

- **Bag of Words (BoW)** - Features using the total number of words of each participant
- **EV** - Features based on eigenvector centrality weighted by degree, excluding IDF influence

The proposed methods to be used as the comparative methods are defined as follows.

- **IDF** - Feature based on IDF
- **AU** - Feature based on discourse marker of argument unit
- **SIM** - Feature based on topic similarity
- **EVIDF** - Feature based on eigenvector centrality on the basis of IDF influence

As a classifier, a Support Vector Machine (SVM) is used since our feature set is relative small to use another classifier (e.g., Random Forest). For SVM kernel, Gaussian kernel is

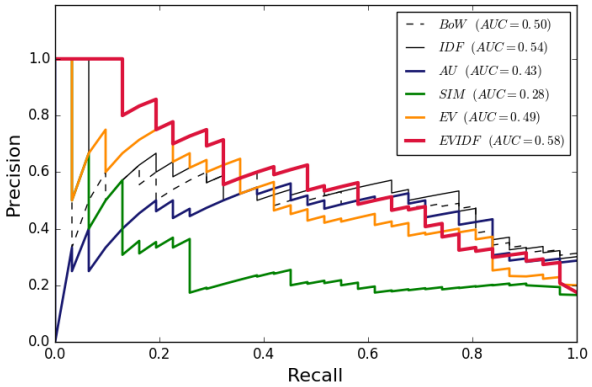

Fig. 3: Precision-recall curve of $\mathcal{H}$

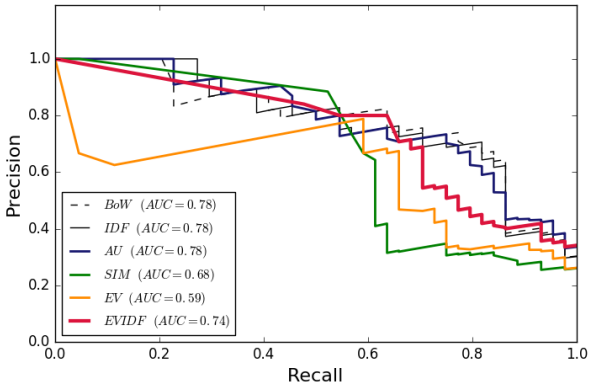

Fig. 4: Precision-recall curve of $\mathcal{L}$

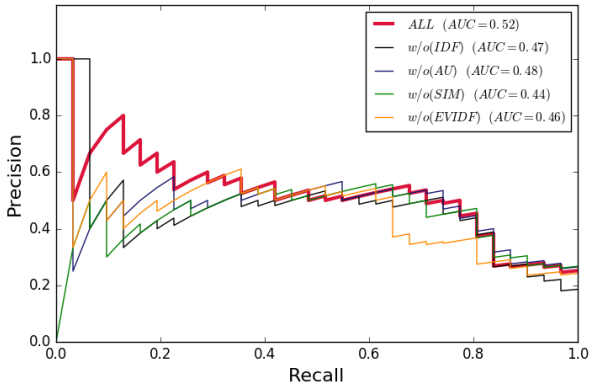

Fig. 5: Precision-recall curve of $\mathcal{H}$ excluding each feature

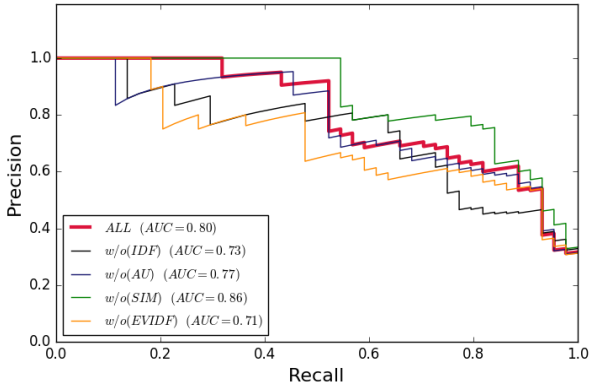

Fig. 6: Precision-recall curve of $\mathcal{L}$ excluding each feature

used. Also, in order to prevent over-fitting, the respective precision-recall curves are the calculated results of the micro-average of each cross-validation using Leave-One-Label-Out (LOGO).

From Fig. 3, it was confirmed that AUCs of IDF and EVIDF models were improved against the baselines, BoW and EV. In particular, since the precision of EVIDF has improved significantly compared with the lexical feature when recall is low, EVIDF is effective in the classification task $\mathcal{H}$. In addition, the model adopting influence is effective since a large difference in AUC is seen between EV and EVIDF. This means that to what extent the lexical feature is spreading to others is having a major influence as judgment criteria of competence under expertise. Of lexical feature quantity, although AU generally shows the same characteristics as those of BoW, its precision is low compared with the baseline. Furthermore, as to SIM, AUC becames low compared with all of the other models. Therefore, we can conclude that the model using AU and SIM does not work well on its own.

According to Fig. 4, AUC of the baseline, BoW, and the proposed methods, IDF and AUC, are the highest. Also, recall and precision show the same tendencies. Furthermore, precision of EV has been lowered significantly and precision of EVIDF has fallen sharply when recall is high. Therefore, for extraction of participants with low expertise as in the classification task $\mathcal{L}$, a simple lexical feature is the most effective. This is because most of the participants who were annotated to have "low" competence in the annotation data are people who seldom make a comment or post very short comments even when they do. In other words, BoW and IDF of simple comments are effective since there is a tendency that participants who make a small amount of comments are evaluated as low due to a lack of judgment criteria.

Evaluation based on combination of features: We will verify the respective effects of the proposed features (IDF, AU, SIM, EVIDF). To do so, precision-recall curve will be compared between the model where all the proposed features are combined and the model where each feature is removed one at a time from the combined model. *ALL* is a model in which all the proposed features are combined. This means that it is a model in which feature quantities of $\{IDF, AU, SIM, EVIDF\}$ are inputs to the classifier. $w/o(f)$ is a model in which the proposed feature f is excluded from *ALL*. For example, $w/o(AU)$ is a model consisting of $\{IDF, SIM, EVIDF\}$.

Fig. 5 and Fig. 6 show the results of the classification task $\mathcal{H}$ and classification task $\mathcal{L}$ respectively. Based on Fig. 5, *ALL*

has the highest AUC. It shows that the model using all of the proposed features is working most effectively. Since AUC of $w/o(SIM)$ is sharply lowered, SIM is working effectively through a combination of other features. Also as to a model excluding EVIDF, it was proved that EVIDF was especially useful as a feature since AUC of $w/o(EVIDF)$ is significantly lowered. Furthermore, as to *ALL*, although it was below IDF and EVIDF when compared with AUC of Fig. 3, a stable curve is drawn with relatively high precision when the recall is high. In Fig. 6, AUC of $w/o(SIM)$ was the highest. As seen above, it was confirmed that topic similarity is not effective as a feature for extraction of a person with lower expertise. However, since AUCs of the other four were lower when compared with *ALL*, features except for topic similarity were useful. In particular, since the decline in accuracy of EVIDF is significant, it is effective to combine the lexical feature and directed influence graph feature. Furthermore, as AUC of all of the features of Fig. 4 is higher than *ALL* and precision of $w/o(SIM)$ is remarkably improved, accuracy can be improved by combining features.

V. Conclusion

This paper focused on the automatic competence estimation of discussion participants' expertise. It proposes the feature extraction method using traditional lexical feature quantity and the directed influence graph feature adopting the influence diffusion model. To be specific, for lexical feature quantity, it proposed the following as features: IDF assuming the difficulty level of words, the number of discourse marker's occurrences focusing on word class, and topic similarity based on TF-IDF cosine similarity. Furthermore, eigenvector centrality of directed influence graph was proposed and new structural feature quantity of the Web discussion bulletin board was defined. As to evaluation experiments, annotated data were created by annotating the expertise of the participants in the data of discussions in the actual Web discussion bulletin board in seven grades. As a result of experiment, it was confirmed that the proposed methods are effective in many cases. As to extraction of persons with high expertise, it confirmed that the feature using argumentative structures works effectively. In addition, as to extraction of persons with low expertise, it was confirmed that the accuracy was improved by combination of the simple lexical feature and directed influence graph feature.

In the future, for the further improvement of accuracy, it is necessary to advance research on the feature extractor. Although we did not conduct experiments on all combinations of features, it will be necessary to evaluate what combinations of features are effective. Furthermore, if the automatic estimation algorithm of expertise will be disclosed to participants, there is a possibility that unfair actions to raise evaluations may be taken. In the future, it wii be necessary to study the robustness with respect to misleading acts against the automatic evaluation algorithm of competence.

References

[1] Takayuki Ito, Y. Imi, M. Sato, Takanori Ito, E. Hideshima. Incentive Mechanism for Managing Large-Scale Internet-Based Discussions on COLLAGREE. in Proceedings of the 3rd Collective Intelligence Conference, 2015.

[2] Takayuki. Ito, Y. Imi, Takanori Ito, E. Hideshima. COLLAGREE: A Faciliator-mediated Large-scale Consensus Support System, in Proceedings of the 2nd Collective Intelligence 2014, 2014.

[3] COLLAGREE, [Online] http://collagree.com/?locale=en

[4] M. Lippi, P. Torroni. Argumentation Mining: State of the Art and Emerging Trends, ACM Transactions on Internet Technology. Vol.16(2): p.10, 2016.

[5] R. Mochales, 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(ICAIL'09), pp.98-107, Association for Computing Machinery, 2009.

[6] E. Barker, R. Gaizauskas, Summarizing multi-party argumentative conversations in reader comment on news. in Proceedings of the 3rd Workshop on Argument Mining, Association for Computational Linguistics, 2016.

[7] S.A. Golder, J. Donath, Social Roles in Electronic Communities. Association of Internet Researchers 5.0, 2004.

[8] J.A. Krosnick, S. Presser. Question and questionnaire design. In P.V. Marsden and J.D. Wright (eds.) Handbook of Survey Research, 2nd edition. Bingley, UK, pp.263-314, 2010.

[9] J. Eckle-Kohler, R. Kluge, 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, 2015.

[10] D.M. Blei, A. Ng, M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, Vol.3, pp.993-1022, January 2003.

[11] D.M. Blei, J.D. Lafferty. Correlated topic models. Advances in neural information processing systems 18, 2005.

[12] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D.M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. Advances in Neural Information Processing Systems 21, 2009.

[13] R. Rinott, L. Dankin, C. Alzate, M. M. Khapra, E. Aharoni, N. Slonim. Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection. in Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing, 2015.

[14] N. Matsumura, Y. Ohsawa, M. Ishizuka. Influence Diffusion Model in Text-Based Communication. Transactions of the Japanese Society for Artificial Intelligence, Vol.17(3), pp.259-267, 2002.

[15] L.C. Freeman, Centrality in networks: I. Conceptual clarification. Social Networks 1, pp.215-239, 1979.

[16] P. Bonacich, Power and centrality: a family of measures. American Journal of Sociology, Vol.92, pp.1170-1182, 1987.

[17] N. Matsumura, Y. Ohsawa, M. Ishizuka. Profiling participants in online?community based on influence diffusion model, Transactions of the Japanese Society for Artificial Intelligence, Vol.18(4), pp.165-172, 2003.

[18] J.L. Fleiss. Measuring nominal scale agreement among many raters. Psychological Bulletin, Vol.76(5), pp.378-382, 1971.

[19] J. Richard Landis, Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. Biometrics, Vol.33(1), pp.159-174, 1977.

[20] J. Davis, M. Goadrich. The Relationship Between PrecisionRecall and ROC Curves. in Proceedings of the 23rd International Conference on Machine Learning (ICML'06), pp.233-240, Association for Computing Machinery, 2006.