Information Engineering Express
International Institute of Applied Informatics
2017, Vol.3, No.4, P.115 - 124

A Visualization Method of Relationships among Topics in a Series of Meetings

Ryotaro Okada, Takafumi Nakanishi*,
Yuichi Tanaka, Yutaka Ogasawara, Kazuhiro Ohashi†,

Abstract

In this paper, we present a visualization method of relationships among topics in a series of meetings. This method is an extension of the previous work: "A Topic Structuration Method for a Meeting from Text Data." The previous work is aimed at analysis of a single meeting, on the other hand, the proposed method is aimed at analysis of multiple meetings. Several meetings that belong to a single project might have common topics. Our visualization method helps us to find these common topics. In addition, the meetings might have isolated topics. Our method also helps to find them. This visualization is useful for review of meetings. We present a preliminary experiment. In the experiment, we show an analysis results of four actual meetings that belong to a single project.

Keywords: Topic Structuration; Efficiency of a meeting; Dialogue; Transition of topics; Time-series data

1 Introduction

In day-to-day production activities, dialogues play very important roles in order to exchange information and create new ideas. In organizations such as companies, many people have some meetings for dialogues to each other. The meeting time occupies a large proportion in business hours. According to the survey by NTT Data Institute of Management [1], in the case of Japanese companies, the time they spend on meetings occupies 15.4% on average in business hours. In addition, the survey asked them to point out problems of meetings. The important problems they mentioned are "There are often unnecessary meetings", "There are too long meetings", and "There are too frequently meetings." We infer from these facts that efficiency improvement of meetings is not enough in spite of meetings occupy much of the working hours.

Generally, improving the efficiency of a meeting means improving the form of a meeting, such as pre-sharing of documents, keeping time, clarification of roles of members, and appointing a facilitator. We should analyze contents and flows of remarks in dialogue on meetings in order to improve efficiency of a meeting.

* Center for Global Communications (GLOCOM) International University of Japan, Tokyo, Japan
† ITOKI Corporation, Tokyo, Japan

In the method of analysis for a meeting, the ethnographic methods [3] have been mainly studied. In these methods, researchers continue to observe the behavior of attendances during the meeting. The methods require the researcher to attend the meeting actually for observation. Generally speaking, it is difficult to keep analyzing by using these methods because of high analysis cost for observing the meeting. Moreover, since the methods are qualitative studies based on observation by human, we cannot practically implement as computer programs at this stage.

On the other hand, it becomes easier to obtain voice data by improvement and cost reduction of the microphone. In addition, it is easy to obtain text data from voice data by improvement of voice recognition technology recently. Based on this background, we design new indexes for efficiency in a meeting by data mining techniques for these text data. An efficiency of a meeting means the elimination of useless meeting, shortening time of meeting, and the promotion of more productive and meaningful meeting.

Facilitating and reviewing a meeting are important for meeting efficiency. It is important to objectively know process of arriving at decision for dialogue by facilitating and reviewing a meeting.

In this paper, we present a visualization method of relationships among topics in a series of meetings. This method is an extension of the previous work: "A Topic Structuration Method for a Meeting from Text Data."[2] The previous work is aimed at analysis of a single meeting, on the other hand, the proposed method is aimed at analysis of multiple meetings.

Several meetings that belong to a single project might have common topics. Our visualization method helps us to find these common topics. In addition, the meetings might have isolated topics. Our method also helps to find them. This is useful for review of meetings.

We define a similarity between topics across meetings, and we implement a visualization method for the similarity. The system visualizes the relationships among the groups in meetings as a graph network based on time series. In the visualization, topic groups are represented as nodes, and similarities among these nodes are represented as edges.

This paper is organized as follows: In section 2, we introduce related works. In section 3, we present our visualization method of multiple meetings. In section 4, we present some preliminary experiments. In section 5, we reach a conclusion.

2 Related Work

There are many factors for analyzing a meeting such as contents they talked, volume of voice, frequency characteristic of voice, attributes of members, behavior of speakers and listeners, etc.

The ethnographic methods [3] consider these factors to evaluate a quality of a meeting. These methods are based on comprehensive observation in dialog analysis. These methods are qualitative evaluation methods and largely depend on the experiences of an observer.

There are also quantitative evaluation methods in dialogue analysis. Researches on analytical methods focusing on non-verbal information such as document (for example, [4][5][6]) are actively conducted. These methods consider non-verbal factors such as the timing of utterance and silence, intonation, pitch / size of voice, speed of speak, etc. to evaluate features of dialogue. These non-verbal information can be easily extracted automatically from voice data.

Our method is the one of analysis methods focusing on language information. In recent years, the performance of speech recognition has much increased due to the development of machine learning technology. It is predicted that various dialogues is accumulated as text data in the near

future. When text mining is applied to these text data, it becomes possible to look back on the dialog effectively and analyze meaning.

One of the best-known methods of extracting topics in text analysis is LDA (latent dirichlet allocation) [7]. LDA is the one of Topic model which is a model to estimate topics using stochastic method. Topics model estimates the latent topics that span multiple documents. In our method, we divide an entire text to segments and extract topics that span multiple segments. The point of extracting topics from multiple parts is common to our method and Topic model. However, our method aims not only to extract topics, but also to review when a topic was spoken, because we focus on improving a meeting process. As studies to divide whole text to parts by topic, Text segmentation [8][9][10][11] methods are usable. Text segmentation methods first divide the text into sentences. Then, the methods measure similarities among these sentences. Finally, the methods identify where the topic switched. As methods of text segmentation, TextTiling[8], C99[9] , etc. are well known. Furthermore, in order to solve the problem of sparseness of word vectors in these methods, studies are also being conducted to compress dimensions using a topic model. There are cases where topic models are applied to TextTiling[10] and applied to C99[11]. The research of text segmentation is in common with our research in that it estimates what and where is spoken in an entire text.

However, in our research, we do not assume that a segment does not necessarily have a common topic with other segments. This is a point of difference from the research of text segmentation. In existing researches on text segmentation have dealt with texts written by human. In contrast, in our research, we target text transcribed from a speech spoken at a meeting. Sometimes, we also have a meaningless conversation in a meeting. A creative process necessarily includes trial and error. Since we focus on the process of a meeting in our research, we aim to analyze not only important conversation but also unimportant conversation.

Until now, we have worked a semantic structuration method on time series for a meeting from text data [2]. In addition, based on this research [2], we have proposed a topic extraction method by the importance which was calculated from the structure of a conversation a meeting [13].

In this paper, based on the research [2], we propose a method of analysis for multiple meetings. We visualize the relationships among the groups in meetings as a graph network. This visualization will provide opportunities for reviewing a meeting from another perspective.

Researches of topic tracking are similar to our research. Dynamic topic model [14] is a typical research of topic tracking. Dynamic topic model is an extended model of topic model [4] to analyze the time series data. In that paper, they extract the temporal transition of topics of scientific journals. In our research, we extract the temporal transition of topics of multiple meetings which are in the same project. we can regard that our research is a topic tracking for meetings.

3 A Visualization Method of Relationships among Topics in a Series of Meetings

In this section, we present a visualization method of relationships among topics in a series of meetings. This method is an extension of the previous work: "A Topic Structuration Method for a Meeting from Text Data."[2] In section 3.1, we present an overview of the previous work. In section 3.2, we present the proposed method. The previous work is aimed at analysis of a single meeting, on the other hand, the proposed method is aimed at analysis of multiple meetings.

3.1 Previous Work: A Topic Structuration Method for a Meeting

In this section, we present an outline of the previous work: "A Topic Structuration Method for a Meeting." Details are shown in the reference [2].

(1) Input data format

In this method, we use text data transcribed by human as input. The format of the text data is "speaker: a sentence", but in this study, we don't use information of speaker. we use only sentences that are spoken. We regard a turnover of speakers as a delimiter of sentences.

(2) Segmentation and normalize

The system divides text data into n segments based on the total number of characters in the entire text. The system divides the text data contained in each segment into words by morphological analysis, and express as a matrix of words and its appearance frequency. In the morphological analysis, we extract only noun words. The system normalizes values of the element in the matrix by TF · iPF method. TF · iPF is the method similar to TF · iDF. TF · iDF is for documents, and TF · iPF is for segments. The details of TF · iPF are shown in the reference [2]. Let TF • iPF value corresponding to each word be a weight of word.

Let n be the number of segments and k be the number of words which occurrence in segments. Let the weight of word j in a segment i be $x_{i,j}$. Let $\boldsymbol{x}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,k}\}$ be a segment vector of segment i. We construct matrix $\boldsymbol{X}$ by arranging segment vectors as rows. $\boldsymbol{X}$ is a matrix representing the weight values of words in all segments.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,k} \end{pmatrix} \in \boldsymbol{R}^{n \times k}$$

(3) Calculation of similarity among segments

The system calculates a similarity value matrix $\boldsymbol{Q}$ from the matrix $\boldsymbol{X}$. Each elements of matrix $\boldsymbol{Q}$ is a cosine similarity of two segments that corresponding to any row and any column in matrix. The dimension of matrix $\boldsymbol{Q}$ is $n \times n$..

$$\boldsymbol{Q} = \{q_{i,j}\} = \begin{pmatrix} q_{1,1} & \cdots & q_{1,n} \\ \vdots & \ddots & \vdots \\ q_{n,1} & \cdots & q_{n,n} \end{pmatrix} \in \boldsymbol{R}^{n \times n}$$

$$q_{i,j} = \cos(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Here, the diagonal component $q_{i,i}$ is 1. Since all of these diagonal components are 1, we ignore them and regard them as 0.

(4) Clustering of Segments and Extraction of Topic Words

The system clusters segments based on similarities among segments. Each segment is expressed as a vector $\boldsymbol{x}_i$. Let ε be threshold for clustering of vectors. When the similarity $q_{i,j}$ between arbitrary two vectors $\boldsymbol{x}_i$ and , $\boldsymbol{x}_j$ is less than ε, the system clusters them. When vectors belonging

to a cluster also belong to other clusters, the system joins those clusters. A cluster has at least two vectors.

Each degree of similarity $q_{i,j}$ between vectors have a value from 0 to 1. When ε is 0, clusters are not created. When ε is 1, one clusters which have all vectors are created.

We think that ε should be set according to purpose. In this research, the system changes ε from 0 to 1, and the system adopts the value which maximize the number of clusters. Generally, the value that satisfies that condition is not one. In this research, the system adopts the smallest value among them.

This method is based on the following three objectives.

- Find low contribution segments: the vectors that are not related to other vectors should not belong to any cluster.
- Discover many topics: a vector which has relation to other vectors should belong to cluster as much as possible.
- Extract topics from each cluster: one cluster should not have too many topics.

Let a cluster of vectors be a group *G*. *G* is a set of vectors. From these groups, we can extract topic words of the meeting. The system picks up the words which commonly exist in the segments belonging to the same group. From these words, the system extracts topic words. A topic word is a word appearing in two or more segments belonging to the same group. Topic words are sorted in descending order by number of vectors which have the topic word. We can extract arbitrary number of topic words from sorted list of topic words.

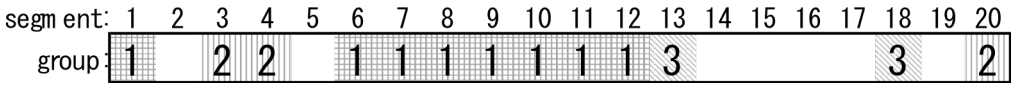
(5) Visualization

In the reference [2], the research shows three types of visualization. Here, we present one of them: “Meeting-Outline”. Meeting-Outline is a visualization for reviewing the flow of the conversation as the overview of the meeting. The system arranges each segment using time series as a coordinate axis. Each segment is classified by color according to the groups. At the same time, the system shows a table of topic words of each group. We can find the rough flow of topics of the meeting by the time series arrangement and coloring.

We show a sample of Meeting-Outline in Figure 1. "****" in the table is a word which we cannot disclose such as a company name, a personal name, etc. In this experiment, we set the number of topic word to 20.

3.2 A Visualization Method of Relationships among Topics in a Series of Meetings

In this section, we present the proposed method: a visualization method of relationships among topics in a series of meetings. This research differs in granularity from the previous research. In the previous research, we aimed to aggregate the transition of topics in one meeting. In this research, we aim to aggregate the transition of topics in multiple meetings which belong to a same project.



1	2	3
region	room	time
position	space	health
****	office	log
life	registration	behavior
area	ICT	comic
attribute	SNS	ability
res	meeting	company
property	provision	
hobby	liking	
concept of values	hobby	
health	device	
downtown	customer	
house	account	
buying and selling	booking	
opulence		
prediction		

Figure 1. Meeting-Outline: Meeting-Outline shows which group each segment belongs to. In addition, the list of topic words shows topic words of each group. In this result, three groups are formed. The group 1 has the most number of segments. Segments 6 to 12 belong to group 1 continuously. We can make a conjecture that the group contains the most important topic. On the other hand, segments 14 to 17 do not belong to any group. We can assume that the discussion was confused there.

3.2.1 Composition of group vectors

We aim to analyze relationships among groups in multiple meetings. In order to make groups comparable, we express one group as one vector. We define a group vector $\boldsymbol{g}_i$ which is a vector representing a group G_i. A group is a set of segments which is expressed as vectors. Group G is defined in section 3.1 (4). Let $\boldsymbol{g}_i$ be an average of vectors belonging to the group G_i. Let the vectors belong to G_i be $\boldsymbol{x} \in G_i$. A group vector $\boldsymbol{g}_i$ is expressed by the following equation. $|G_i|$ shows the number of elements of set G_i.

$$\boldsymbol{g}_i = \frac{\sum_{\boldsymbol{x} \in G_i} \boldsymbol{x}}{|G_i|}$$

3.2.2 Visualization: Graph network on similarity

There are multiple meetings within a project. Let one meeting be M_i, let the next meeting be M_{i+1}. The system computes all combination of similarities among the group vectors belonged to M_i and belonged to M_{i+1}. We use a cosine measure for the similarity calculation. In this case, the

group vectors in different meetings correspond to different word sets. Therefore, the system reconstructs group vectors based on the union of the word sets appearing two groups corresponding to two meetings. The elements of vectors that corresponding to the words that did not exist in own meeting be 0.

The similarity $z_{i,a,b}$, between group a in meeting M_i and group b in meeting M_{i+1} is defined as the following equation. $\boldsymbol{g}_{i,a}$ is the group vector of group a in meeting M_i, and $\boldsymbol{g}_{i+1,b}$ is the group vector of group b in meeting M_{i+1}. Function *cos* shows the similarity calculation by the cosine measure.

$$z_{i,a,b} = cos(\boldsymbol{g}_{i,a}, \boldsymbol{g}_{i+1,b})$$

The system visualizes the relationships among the groups in meetings as a graph network based on time series. First, the system visualizes each meeting individually. The system shows each meeting as a set of groups that have several topic words. In this graph, a group corresponds to a node. Second, the system puts these meetings from left to right based on time series. Finally, the system draws edges among nodes based on value of z. When $0 < z <= 1/3$, the system doesn't draw a line. When $1/3 < z <= 2/3$, the system describes a thin line. When $2/3 < z <= 1$, the system draws a bold line. We show an example in the section of experiment.

These criteria for drawing lines are temporary. In the future work, we aim to enable interactive drawing after the first drawing. In this paper, we set the criteria so that the three cases appear almost equally, based on the experimental results.

4 Preliminary Experiment

In this section, we present a preliminary experiment. We implemented our proposed method. We apply it to an actual series of meetings, and show the example of visualization.

4.1 Experiment Environment

We implement the experiment system by Python.

In this experiment, we use text data of actual meetings transcribed by human. For our future work, we will apply automatic voice recognition. The language of those texts is Japanese. We analyze 4 meetings. We analyze a series of meetings within one project. We analyze four meetings. Those meetings were conducted by different days. The members of the meetings were mostly fixed, but not perfectly. Table 1 shows the number of characters (in Japanese) and the total time of each meeting as overview of these four meetings.

Table 1. Overview of meetings for experiment

	Meeting-1	Meeting-2	Meeting-3	Meeting-4
Number of characters (Japanese)	13,826	20,106	16,748	12,622
Total time	2h18m	2h02m	1h55m	1h44m

In this experiment, we show an analysis results of these meetings.

4.2 Experimental Result

Figure 2 shows the visualization of the proposed method in the experiment.

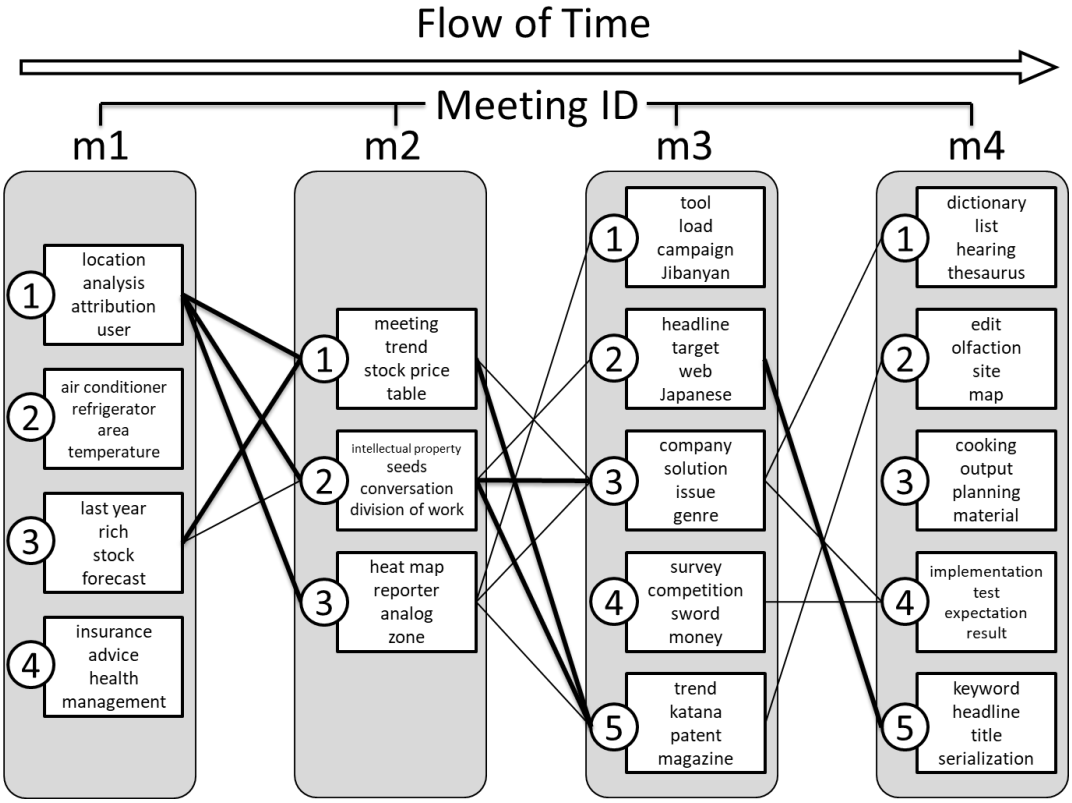


Figure 2. The visualization of relationships among topics in a series of meetings. There are four meetings, and the meetings has from 3 to 5 topic groups. These groups are connected by edges which shows similarity between a node and another node. The thickness of the line represents the strength of the relation. Each topic of group is represented by four words.

In this visualization, each group in each meeting is shown as a node. Each meeting doesn't necessarily have the same number of groups. In this figure, we show four words for representing each topic of groups in each meeting. These four words are the top four topic words of each group. The number of topic words to be displayed is not limited to four, and it will be necessary to adjust according to the purpose. Finally, we draw edges based on similarity among groups.

We will consider the results. From the Figure 2, we can find that there are topics that are not connect to other topics of following meetings. In this case, there are group 2 and group 4 in meeting 1 and group 1 in meeting 3. These topic words actually have not appeared in topic words of other meetings. We can understand that these topics have weak relationships to other topics. We can consider these groups as at least two interpretations: the groups were useless or were ignored simply. Here, we can't judge them. In any case, this visualization seems to be effective for looking back on topics in the project.

Next, we focus on edges. By looking at a pair of nodes connected by bold edges, we can find multiple examples where the same word appears. Examples are "stock" in meeting1-group3 and meeting2-group1, and "headline" in meeting3-group2 and meeting4-group5. It shows that we

can follow topics across the meetings by focusing on edges. Even when the same word does not appear, we can find a case where a pair of groups has relation. An example is meeting1-group1 and meeting2-group3. At there, topic words of meeting1-group1 are "location, analysis, attribution, user", and meeting2-group3 are "heat map, reporter, analog, zone". From the words such as "location, analysis, attribution, user, heat map, zone", we can guess the following things: participants in the meetings talked about analyzing customer information based on location information and mapping it on a map for visualization. Even when we cannot find relation at a glance, edges help us to find a relation between groups by looking at the details along the edge.

We can think that the topics which span multiple meetings are important topics relevant to the entire project. Therefore, finding such topics is helpful for review of the project. We can also think that nodes with many edges have many topics related to the entire project. By looking at the details of the nodes in descending order from the nodes which has many edges, reviewers can look back on the project efficiently. Conversely, we can think that the topics in a node without edges are not topics related to the entire project. Therefore, we can know the importance of a node by checking the number of connected edges. We can think that the node which has many edges is important, and the node has no edge is less important. By analyzing those node in detail, we may find knowledge that contributes to productivity improvement.

5 Conclusion

We presented a visualization method of relationships among topics in a series of meetings. This method is an extension of the previous work: "A Topic Structuration Method for a Meeting from Text Data." The previous work is aimed at analysis of a single meeting, on the other hand, the proposed method is aimed at analysis of multiple meetings.

We defined a similarity between topics across meetings, and we implemented a visualization method for the similarity. The system visualizes the relationships among the groups in meetings as a graph network based on time series. In the visualization, topic groups are represented as nodes, and similarities among these nodes are represented as edges.

Moreover, we presented a preliminary experiment. In the experiment, we showed an analysis results of four meetings that belong to a single project. From the results, we were able to find some knowledge from the visualization.

References

[1] NTT DATA Institute of Management Consulting, “Survey on 'Conference innovation and work style' (in Japanese),” https://www.keieiken.co.jp/aboutus/newsrelease/121005/; retr. 2017/3/23.

[2] R. Okada, T. Nakanishi, Y. Tanaka, Y. Ogasawara, K. Ohashi, “A Topic Structuration Method on Time Series for a Meeting from Text Data,” In: Lee R. (eds) Studies in Computational Intelligence, Springer, Cham, Vol. 721, pp.45-59, 2017.

[3] D. Cameron, Working With Spoken Discouse, SAGE Publications Ltd., 2001.

[4] J.M. DiMicco, K.J. Hollenbach, A. Pandolfo, W.Bender, "The Impact of Increased Awareness While Face-to-Face," Human–Computer Interaction, Vol. 22(1-2), pp.47-96, 2007.

[5] T. Bergstrom and K. Karahalios, "Conversation Clock: Visualizing audio patterns in co-located groups," In Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS '07). IEEE Computer Society, Washington, DC, USA, page 78, 2007.

[6] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) vol.39(1), pp.43-55, 2009.

[7] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation", in Journal of Machine Learning Research, pp. 1107-1135, 2003.

[8] M. A. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages." Computational linguistics, vol. 23(1), pp. 33-64, 1997.

[9] F. Y. Choi, "Advances in domain independent linear text segmentation." Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. Association for Computational Linguistics, pp.26-33,2000.

[10] M. Riedl and C. Biemann, "How text segmentation algorithms gain from topic models," In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 553-557, 2012.

[11] M. Riedl and C. Biemann, "TopicTiling: a text segmentation algorithm based on LDA," In Proceedings of ACL 2012 Student Research Workshop (ACL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, pp.37-42, 2012.

[12] R. A. Baeza-Yates and B. A. Ribeiro-Neto. Modern information retrieval: the concepts and technology behind Search (2nd Edition). Addison-Wesley Professional, 2011.

[13] T. Nakanishi, R. Okada, Y. Tanaka, Y. Ogasawara, K. Ohashi, "A Topic Extraction Method on the Flow of Conversation in Meetings," In proceedings of 6th International Congress on Advanced Applied Informatics (AAI2017), pp.350-355, 2017.

[14] D. Blei, A. Ng, and M. Jordan, "Dynamic topic models", Proceedings of the 23rd international conference on Machine learning, ACM, pp.113-120, 2006.