Analysis on the Usage of Topic Model with Background Knowledge inside Discussion Activity in Industrial Engineering Context

Muhammad Luthfi*, Satoshi Goto† and Osamu Yoshie‡
Graduate School of Information, Production, and Systems
Waseda University, Kitakyushu, Japan 808-0135
**muhammad.luthfi@akane.waseda.jp*, †*satoshi.goto@fuji.waseda.jp*, ‡*yoshie@waseda.jp*

***Abstract*—Consensus building process for enterprise digital transformation is a significant approach on the implementation of Internet of Things (IoT) solutions through product lifecycle management (PLM). When we improve the consensus building process, it is important to find any latent opinions and hidden dialog patterns analyzing discussion activities by stakeholders. Several approaches have been proposed in forms of instructions and frameworks such as causal model of Consensus Building Theory (CBT) and short-term intensive workshop in strategy planning phase of Product Lifecycle Management (PLM) process. This paper will analyze a new approach to improve consensus building process by summarizing discussion activity. The proposed method is done by performing data augmentation and topic modeling with the help of background knowledge on discussion activity held within industrial engineering context. Our method produces a complete summarization of discussion activity that consists of topic distribution and distribution similarity between topics. We also found that the usage of data augmentation and background knowledge will improve topic quality. We validate our findings to a professional consultant and conclude that our approach gives an adequate contribution towards summarizing discussion activity that might improve consensus building process.**

***Keywords*-topic model; background knowledge; consensus building; product lifecycle management; data augmentation**

I. Introduction

A conventional discussion activity happened when a group of people let out their own opinion with appropriate feedbacks from the other. In industries, discussions are being held in various departments to solve specific problems. We can characterize such discussions as a group of people who shares a same interest aimed to build one single consensus.

Some important features of consensus building process are recording, facilitation, and mediation [1]. Recording in this term stands for creating a physical record of what subject being discussed. Recording can be implemented by recording the whole discussion as a video file or even as simple as taking notes on participant's utterance. Facilitation in a second hand, help participants work together by providing artefact containing the discussion progress which everyone agrees on. Finally, mediation acts to help opposite parties deal with disagreement. In order to perform mediation, one independent person is needed to resolve disputes with his/her objective point of view.

Consensus building plays a significant role in strategy planning phase of Product Lifecyle Management (PLM) process [2]. Strategy planning phase is involving people from various levels and departments in an organization [1], thus justifies the needs of consensus building approach. The practice of consensus building often still have frequent problems. During discussion activities, various stakeholders with different personalities and backgrounds might influence conclusion [3] which will affect tendency and direction of the discussion [4].

PLM practice also supports manufacturing companies towards digital transformation [5]. As one of the implementations of Internet of Things (IoT) solutions, PLM with its holistic paradigm helped companies to change their internal resources e.g. business processes, product data, and people by taking the benefits of its external resource generated by other IoT solutions. In this research, we propose a better consensus building approach for PLM process which will leads to the advancement of digital transformation.

II. Related Works

Researches related to consensus building have been conducted years ago. In general, researches focused on consensus building can be divided into 2 categories based on their focus point; process model and measurement model. All models are designed under the same goal: to improve consensus quality.

Process model focused on a set of rules that participants should follow under specific circumstances. A research proposed a straightforward approach using a fair, open, and freedom-focused process model [8], meaning that all perspective will be considered equal and all participants will have their freedom to disagree. Another research is focused specifically on a subprocess in consensus building like Consensus Building Theory (CBT) [9] which emphasizes the cause of conflict to investigate what specific matter prevent or support consensus building. Other research is focused on a specific implementation of process model [4] e.g. proposing a short term and intensive workshop activity designed for Product Lifecycle Management (PLM) strategy planning phase which involves multi-party stakeholders. The

workshop is intended specifically for discussion under digital transformation for smart, connected engineering field.

Meanwhile, measurement model focused more on the criteria to determine consensus presence. Some popular methods are done by using standard deviation of voting results or using Kendall's coefficient on voting results [10]. A recent method showed that a digitized approach can be done by tracking every non-verbal aspects of each participant to determine consensus [7]. Another digitized approach has been done in 2 steps: inviting external facilitator as one independent figure to direct the course of discussion and doing text mining approach on the dialog data taken from the discussion to evaluate consensus [6].

III. Research Problem

Previous researches discussed on how consensus is being built by considering a lot of variables e.g. time consumed, participants contributions, and conflict resolution method. In another hand, more variables should be valued more such as the objectivity of the final consensus and the latent opinion from each participant. A preliminary study regarding this matter has been conducted [6] but it risks of having biased judgment which will damage consensus building process. We proposed a better text mining approach by utilizing topic model algorithms, a set of data as Background Knowledge, and calculating the convergence rate among topic distributions. Data augmentation is also added in the process to increase data quality.

IV. Proposed Method

In this research, we performed a digitized approach of dialog data from PLM-themed discussion activity sessions using data augmentation, topic model with background knowledge, and distribution similarity. First, the data will be prepared by a simple preprocess method and data augmentation. The clean and augmented data will then be experimented by various topic models and hyperparameters, we picked the best configuration and incorporate it into background-knowledge-backed topic model to generate topic distributions. Then, we will calculate the distribution similarity as convergence rate. Finally, we will compare the effectiveness of our approach to previous research and get a professional consultant to analyze the topic distribution results to have an objective review. To summarize, we will take dialog data of discussion session and transform it into topic distributions, similarity value, and most frequent words (if necessary) from each discussion session to be validated by a professional consultant.

A. Data Augmentation

We took a real-life dialog data from discussion sessions which ran for 1-2 hour long. Based on the dataset characteristics in Table I, the dataset we used is very poor. Compared to the common dataset in topic model research with specialization in short text data, our dataset size is 96.55% lower in terms of number of documents and 85.96% lower in terms of corpus size. Hence, we are using data augmentation techniques to improve dataset quality. We expand the Easy Data Augmentation [12] by adding additional processes: hypernym replacement and hyponym replacement. Hypernym and hyponym of a word is crucial as we thought the topic mixture of a sentence *s* should be the same with another sentence *s'* who has hypernym/hyponym relation with it.

Table I
Dataset Characteristics

Measures Type	PLM Workshop Dataset	Common Dataset [11]
Total Documents	383	11094
Corpus Size	686	4887
Average Length	4.83	7.84

B. Topic Model with Background Knowledge

We tried to mine latent opinion of the dataset using topic model with background knowledge. Topic model is an unsupervised learning approach where we could transform documents into document-to-topic distributions and topic-to-word distributions. In topic model point of view, document is a mixture of topic where topic itself is a mixture of word. The most popular method of topic model is Latent Dirichlet Allocation (LDA) [13], in which, current topic model researches mostly use LDA as baseline method. In LDA-based topic model, the learning process consists of generation process and sampling process. In generative process, the initial document-to-topic distributions and topic-to-word distributions are generated using hyperparameter α and β. Then, in the sampling process, distributions are evaluated by recalculating it using Gibbs Sampling for each word. The graphical notation of LDA topic model is shown in Fig. 1. Meanwhile, the generation algorithm is shown in Algorithm 1 where *K* is number of topics, *D* is number of documents, N_d is number of words in document d, ϕ_k is topic-to-word distribution for topic k, θ_d is document-to-

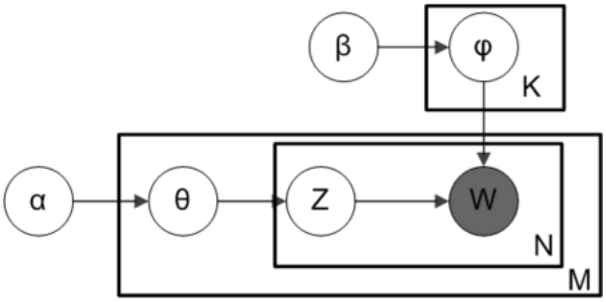


Figure 1. LDA plate notation

Algorithm 1 Generation Algorithm of LDA

for $k \in \{1, ..., K\}$ **do**
  Generate $\phi_k \sim$ Dir(β)
for $d \in \{1, ..., D\}$ **do**
  Generate $\theta_d \sim$ Dir(α)
for i, d where $d \in \{1, ..., D\}$ and $i \in \{1, ..., N_d\}$ **do**
  Generate $z_{id} \sim$ Multinomial(θ_d)
  Generate $w_{id} \sim$ Multinomial($\phi_{z_{id}}$)

topic distribution for document *d*, z_{id} is topic for the i_{th} word in document *d*, and w_{id} is the i_{th} word in document *d*.

We mentioned in previous paragraph that our dataset has relatively smaller size compared to common topic model researches. Hence, we assembled various topic models with specialty in short text as suggested by [11]. The whole list of topic models could be seen in Table II that could be categorized into 4 types. The first type is standard, referred to the baseline topic model which is LDA. The second type, one-topic sampling based, will modify the inference process to sample only one topic per document, meaning all words from a single document will only have one topic label. The third type, global word co-occurrence based, will modify document representation into word-network or set of biterms. Self-aggregation based as the last type will merge several documents into one single pseudo-document and then apply standard-type topic model to it.

After the experiment is done, we will decide what is the best topic model, hyperparameters, and the number of sentence augmentation processes to use. After that, we will incorporate the result to a new background-knowledge-backed topic model called Source-LDA [14] as the most suitable topic model for our case. In Source-LDA, we can use background knowledge data to influence topic labeling thus improving topic quality in the process.

Table II
TOPIC MODEL EXPERIMENT

No.	Topic Model	Type
1	LDA [13]	Standard
2	Dirichlet Multinomial Mixture (DMM) [16]	One-topic sampling based
3	Latent-Feature LDA (LFLDA) [17]	
4	Latent-Feature DMM (LFDMM) [17]	
5	Generalized Polya Urn DMM (GPU-DMM) [18]	
6	GPU Poisson-based DMM (GPU-PDMM) [19]	
7	Biterm Topic Model (BTM) [20]	Global word co-occurence based
8	Word Network Topic Model (WNTM) [21]	
9	Self-aggregate Topic Model (SATM) [22]	Self-aggregation based
10	Pseudo-based Topic Model (PTM) [23]	

C. Distribution Similarity

In this step, we aimed to picture the topic distribution into a single value that describes the rate of consensus built (convergence rate). In order to do this, we used distribution similarity calculation using Jensen-Shannon Divergence across all distributions [15]. This concludes the final step of our proposed method.

V. EXPERIMENT

We had an opportunity to utilize dialog data from requirement decisions (discussion session) of 4 Japanese companies. Firstly, data preprocessing and sentence augmentation is done to clean the data. The comparison of dataset characteristics before and after augmentation is shown in Table III. We managed to expand the dataset up to 1634.46% from the original size in terms of number of documents, and up to 145.04% in terms of corpus size.

Furthermore, the dataset property is presented in Table IV. During dialog data collecting process, 2 types of question were asked. Problem-type question was given at the early stage of discussion while solution-type question is asked at the later stage of discussion. Another property is 'response category' that referred to participant's own division, and 'organization level' that referred to participant's hierarchical level in the company.

Following data preprocessing step, topic model experiment is conducted on all topic models in Table II. We used

Table III
AUGMENTED DATASET CHARACTERISTICS

Augmentation	Total Documents	Corpus Size	Average Length
No Augmentation	383	686	4.83
1 Sentence Augmentation	1017	922	5.00
9 Sentence Augmentation	5085	1519	5.10
12 Sentence Augmentation	6643	1681	5.10

Table IV
DATASET PROPERTY

Property Name	Possible value
Company ID	{1,2,3,4}
Question Type	{Problem, Solution}
Response Category	{Information Technology, Corporate Management, Business Process, Human Development}
Organization Level	{very low, low, medium, high, very high}
Opinion	{short sentence consists around 5 words}

topic coherence value to evaluate topic model performance because our dataset is raw and doesn't have any golden labels [11]. The result of topic model experiment is shown in Fig. 2. The hyperparameter used in this experiment is number of iteration (1000-2000), α value (0.05-0.3), and β value (0.005-0.03). The number shown in the figure is the average of topic coherence value of all possible hyperparameters for each sentence augmentation processes. Based on the number of sentence augmentation, 9 augmentation is not producing significant result while 1 augmentation gives the best and most consistent result. Self-Augment Topic Model (SATM) gives a good overall score regardless the number of sentence augmentation process. However, LDA held the best score of all experiment with 1 augmentation process. We picked LDA topic model with α value of 0.15, β value of 0.01, 2000 iteration, and 1 sentence augmentation process as the best configuration.

The next step is to incorporate the best configuration into Source-LDA. Based on our proposed method in Section IV, we are using PLM-themed discussion activity as our dataset. Hence, PLM topics is used as the background knowledge data. We decided to use PTC Value Roadmap[1] because it contains many PLM Topics with complete definitions for each topic. The background knowledge dataset held a relatively big size consisting of 26 topics, 1068 unique words, and 145.88 average document length.

Fig. 3 shows the topic coherence value relative to the number of sentence augmentation process applied to background knowledge dataset. The topic coherence value is increasing proportionally with the number of sentence augmentation. We only used LDA algorithm because it is the best configuration we conclude from previous paragraph. The usage of sentence augmentation on background knowledge improves topic coherence up to 3.25%. We picked 12 sentence augmentation process on background knowledge as the best configuration.

[1]http://support.ptc.com/WCMS/files/28837/en/J1051_ValueRoadmap_TS.pdf

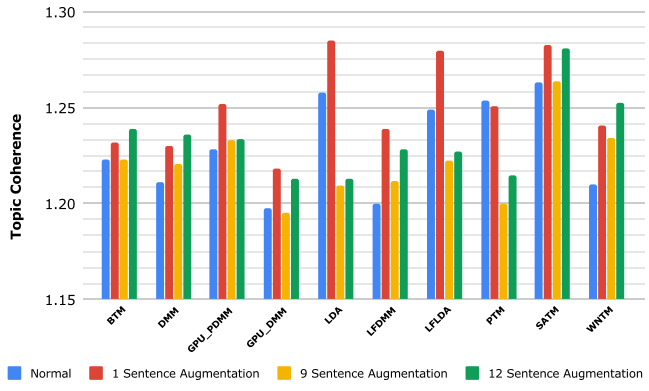


Figure 2. Topic coherence value based on topic model and sentence augmentation process performed on Dialog Data

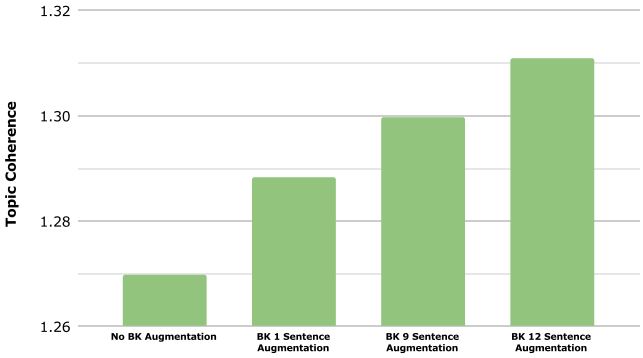


Figure 3. Topic coherence value based on background knowledge (BK) augmentation process

The final best configuration we found is by using LDA topic model, α value of 0.15, β value of 0.01, 2000 iteration, 1 sentence augmentation on dialog dataset, and 12 sentence augmentation on background knowledge. We are planning to use this finding in future experiment.

VI. Results and Discussion

In this section, we compared our findings with previous research [6] as shown in Fig. 4. We compared between both approaches using topic coherence value. We improved the result by 6.34% by utilizing data augmentation and background knowledge, .

Furthermore, we applied our model to create topic distributions of the dataset. Then, we calculate the convergence rate among topic distributions and do qualitative evaluation together with external facilitator.

During topic modeling process, we replaced background knowledge's topic name with its topic number to increase

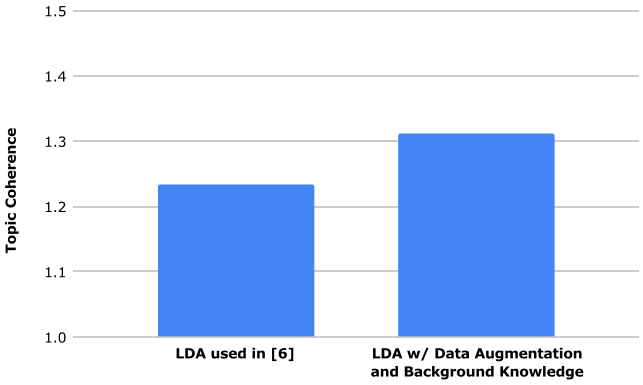


Figure 4. Topic Coherence value comparison with previous research

readability. The mapping of topic number with its actual name can be seen at Table V with a side note that the order of topic number is in alphabetical order and different with what shown in the reference (PTC Value Roadmap). There are 26 PLM Topics in total that serves as best practice for a specific business unit.

The average topic distribution of sentences in dialog data from each company is shown in Fig. 5. The most probable and least probable topic is different for each company. For example, company 1 that runs in business process innovation industry discussed a lot about topic #12 (Project Management) and very little about topic #24 (Verification and Validation). Meanwhile, Company 2 from automotive manufacturer industry had a huge interest in topic #18 (Service Order Management and Field Service) but not in topic #22 (Technical and Service Parts Information Creation and Delivery). Company 3 from aqua industry has topic #19 (Service Parts Planning and Pricing) as the most probable topic and topic #2 (Component and Supplier Management) as the least probable topic. Lastly, company 4 from automotive supplier industry had topic #22 and topic #21 (System Architecture Design) as their most probable topics.

Table V
MAPPING OF PLM TOPICS

Topic No.	PLM Topics
Topic 0	Business System Support
Topic 1	Change and Configuration Management
Topic 2	Component and Supplier Management
Topic 3	Concept Development
Topic 4	Design and Manufacturing Outsourcing
Topic 5	Equipment Monitoring and Lifecycle Management
Topic 6	Manufacturing Process Management
Topic 7	Mechanical, Electrical, and Software Development
Topic 8	Performance Analysis and Feedback
Topic 9	Platform Design and Variant Generation
Topic 10	Product Cost Management
Topic 11	Product Support Analysis and Planning
Topic 12	Project Management
Topic 13	Quality and Reliability Management
Topic 14	Regulatory and Materials Compliance
Topic 15	Requirements Definition and Management
Topic 16	Service Diagnostics and Knowledge Management
Topic 17	Service Logistics and Network Management
Topic 18	Service Order Management and Field Service
Topic 19	Service Parts Planning and Pricing
Topic 20	Smart, Connected Product Enablement
Topic 21	System Architecture Design
Topic 22	Technical and Service Parts Information Creation and Delivery
Topic 23	Tooling Design and Manufacture
Topic 24	Verification and Validation
Topic 25	Warranty and Performance-based Contract Management

Given the topic distributions, we will conduct similarity measurement using JS Divergence to find convergence rate. The interpretation value of JS divergence ranges from 0-1. All value approaching to 0 means that there is no variation between probability distributions, meanwhile value approaching to 1 means that there is high variation between probability distributions. The convergence of each discussion sessions can be seen at Table VI along with the top frequent words. The overall convergence rate achieved from each discussion is probably not too good since each one has convergence rate above 0.500. The lowest degree of convergence was achieved by Company 1 with 0.865 while the highest degree of convergence was achieved by Company 3 with 0.672. The top words from each discussion acts as a support to better understand the discussion.

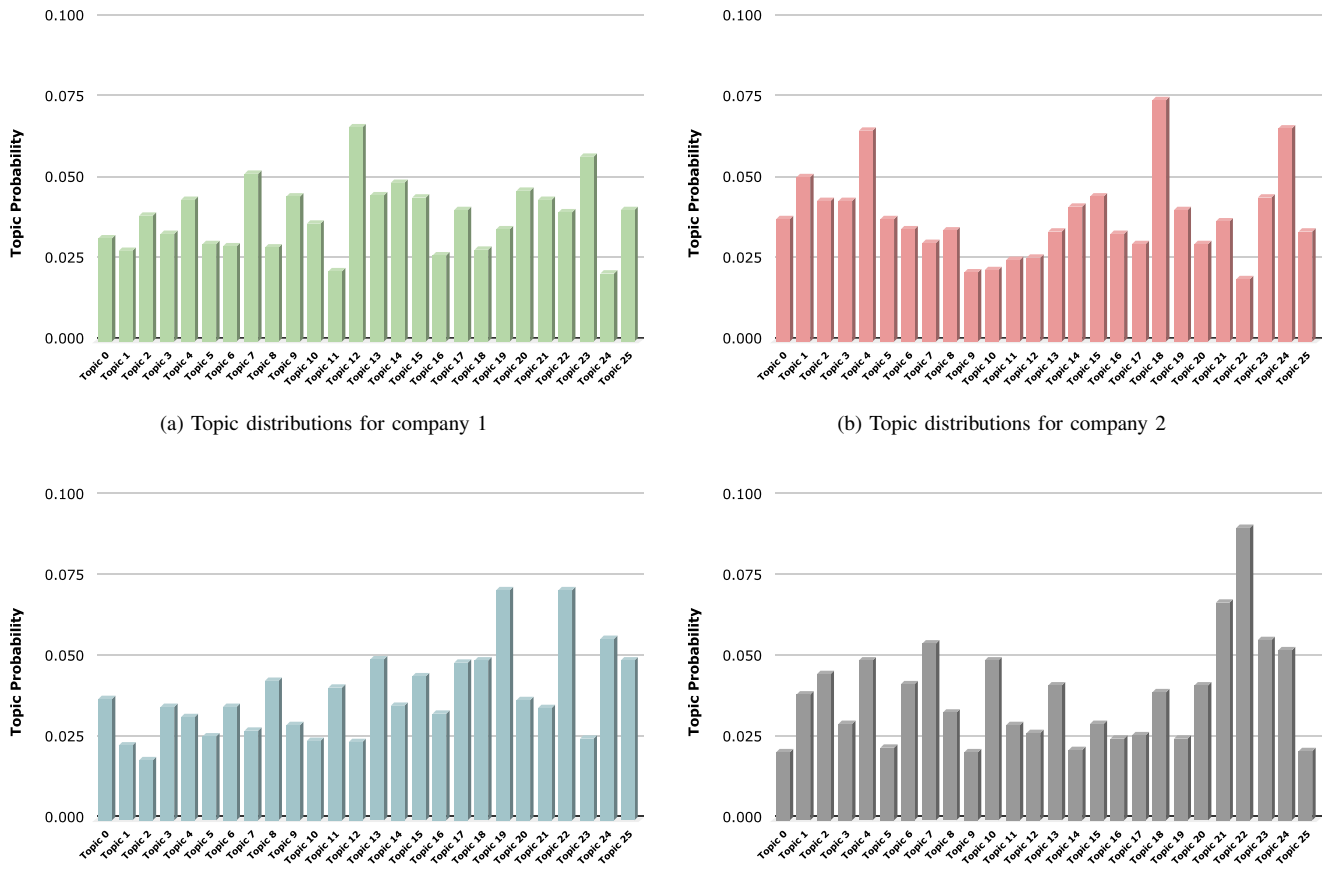
Given the results, here is the feedback from professional consultant for each discussion session:

1) Company 1: Company 1 had a key problem in terms of information exchange between design and manufacturing. I agree that the frequency of Design and Manufacturing topics was high. However, the topic of Project Management was rarely spoken directly by their voices. In addition, the analysis results show that there are few topics on Manufacturing Process Management. Certainly, there were few remarks on Manufacturing Process Management when the workshop was actually held. However, one of the participants was very concerned about the topic and he is one of the important people in the PLM project, so even if it is a minority opinions, I cannot ignore it as my consultant perspective. By the way, in the analysis results, the words with the highest frequency of occurrence were Information, Product, and Data. These were key words that participants often talked about during the actual workshop. As a consultant, I agree with that.

2) Company 2: The company 2 had three business unit. Thus, the participants had different opinions, as each business unit had a completely different product and each business model was different. When I looked at the results

Table VI
CONVERGENCE RATE AND TOP WORDS

Company ID	Convergence Rate	Top words
Company 1	0.865	{Information, Product, Data}
Company 2	0.766	{Production, Work, Product}
Company 3	0.672	{Resource, Human, Product, Development}
Company 4	0.753	{Information, Data, Sharing}



(a) Topic distributions for company 1

(b) Topic distributions for company 2

(c) Topic distributions for company 3

(d) Topic distributions for company 4

Figure 5. Average topic distributions of sentences in dialog data of all company

of this analysis, I thought that the reason that the topic of Verification and Validation was high was probably that they had a problem with their product quality. However, although the topic about Field Service has not been talked about in the actual workshop time, the frequency of topic 18 was high in this analysis result. In fact, this company does little field service work, so it is necessary to confirm why such analysis results were performed. In addition, the analysis results indicated that the frequency of Product Cost Management and Project Management topics was low. However, I think the discussions about costs and projects were relatively common during the actual discussions with them. Regarding the word distribution, the analysis result, it showed that the frequency of Production, Work, and Product were high. I agree with this result.

3) Company 3: The motivation for Company 3 to introduce PLM was to strengthen its field service operations. Looking at the analysis results, it was found that the topics with the highest frequency were field services, such as Warranty management, Performance Based Contract Management, Technical Service Parts Information, and Service Order Management. I agree this result as a professional consultant. However, regarding the monitoring and management of equipment, it was analyzed that the topic frequency was low. This is different from the actual situation, because in the actual workshop, the story of equipment monitoring was relatively well discussed. The frequency of words of Resource, Human, Product, and Development is high. Even during the actual workshop discussion, the shortage of human resources in field service was very problematic. Thus, I agree with the analysis results.

4) Company 4: Company 4 has been practicing efforts to make its factory a smart factory. As a consultant, what I noticed in their actual workshops was their lack of information sharing between departments and insufficient training of employees. On the other hand, looking at the results of this analysis, we found that the topic # 22 was Technical and Service Part Information Creation and Delivery. At first, I wasn't interested in topic # 22. However, after reviewing the content of discussions with the workshop participants at a later date, there was an opinion that attention was paid to the management of service parts in order to contribute to

sustainable sales. It seems that the results of this analysis have taught me a topic that I did not notice at first. Looking at the analysis results of the word distribution, it seems that three words, Information, Data, and Sharing, appear frequently. This was exactly the issue that was being talked about at the workshop. Additionally, The analysis results seem to indicate that there is no relationship between education and system design. Further investigation is needed as I think education topic should be highly related in the workshop.

Based on our analysis and feedback from professional consultant. We feel that our experiment on the usage of topic model with background knowledge in an industrial engineering discussion activity (in this case, PLM-themed) gives an actual contribution towards discussion summarization in which, might improve consensus building process. The important takeaway of this research is that topic modeling with background knowledge will assist professional consultant to understand more towards participant's latent opinion.

VII. Conclusion and Future Works

In this paper, we analyzed a new digitized approach to improve consensus building process in discussion activity held within industrial engineering context (PLM-themed). Our proposed method consists of performing data augmentation, implementing topic model with background knowledge, and calculating the distribution similarity. Finally, we validate the result on professional consultant. We received good feedback which validates our purpose of using a new approach to improve consensus building process. Moreover, we also found that using data augmentation and background knowledge in topic modeling will improve its topic quality.

However, further approach is still necessary based on two perspective: consensus building and topic modeling. From consensus building perspective, we still need to assure the emotional state of discussion participants when dialog data is recorded. Some variables might aspect the quality and consistency of participant's opinion. Meanwhile from topic modeling perspective, we are planning to expand Source-LDA so it can afford different data representation like BTM and WNTM does.

References

[1] J. Thomas-Lamar, S. Mckearnan, and L. Susskind, *The consensus building handbook: a comprehensive guide to reaching agreement*. SAGE Publications, 1999. pp. 7–9.

[2] J. Stark, *Product lifecycle management*. Springer, 2015.

[3] N. He, S. Yao, and O. Yoshie, *Emotional speech classification in consensus building*. 2014 10th International Conference on Communications (COMM), Bucharest, 2014, pp. 1-4.

[4] S. Goto, O. Yoshie, and S. Fujimura, *Preliminary Study: Text mining approach to dialog data of stakeholders on requirement decision for Enterprise Information System*. 2019 10th Annual European Decision Sciences Institute (EDSI) Conference, Nottingham, 2019.

[5] C. H. Matt and A. Benlian, *Digital transformation strategies*. Business & Information Systems Engineering Vol. 57, 2015, pp. 339-343.

[6] S. Goto, O. Yoshie, and S. Fujimura, *Empirical study of multi-party workshop facilitation in strategy planning phase for Product Lifecycle Management (PLM) system*. 2019 International Federation for Information Processing (IFIP) International Conference on Product Lifecycle Management, Moscow, 2019, pp. 82-93.

[7] Y. Katagiri et al., *Implicit proposal filtering in multi-party consensus-building conversations*. Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus, 2008, pp. 100-103.

[8] C. T. L. Butler and A. Rothstein, *On conflict and consensus; A handbook on formal consesus decision making*. Food Not Bombs Publishing, Takoma Park, 2004.

[9] R. O. Briggs, G. Kolfschoten, and G. J. de Vreede, *Toward a theoretical model of consensus building*. 11th Americas Conference on Information Systems, 2005.

[10] M. M. Shepherd and W. B. Martz Jr, *Group consensus: Do we know it when we see it?* Proceedings of the Hawaii International Conference on System Science, Los Alamitos, 2004.

[11] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, *Short text topic modeling techniques, applications, and performance: a survey*. 2019.

[12] J. Wei and K. Zou, *EDA: Easy Data Augmentation techniques for boosting performance on text classification tasks*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, 2019, pp. 6383-6389.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent Dirichlet Allocation*. Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, Vancouver, 2001, pp. 601-608.

[14] J. Wood, P. Tan, W. Wang, and C. Arnold, *Source-LDA: Enhancing probabilistic topic models using prior knowledge sources*. 33rd IEEE International Conference on Data Engineering (ICDE), San Diego, 2017, pp. 411-422.

[15] J. A. Aslam and V. Pavlu, *Query hardness estimation using Jensen-Shannon Divergence among multiple scoring functions*. 29th European Conference on Information Retrieval Research (ECIR), Rome, 2007, pp. 198-209.

[16] J. Yin and J. Wang, *A dirichlet multinomial mixture model-based approach for short text clustering*. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York,2014, pp. 233-242.

[17] D. Q. Nyugen, R. Billingsley, L. Du, and M. Johnson, *Improving topic models with latent feature word representations*. Transactions of the Association for Computational Linguistics vol. 3, 2015, pp. 299-313.

[18] C. Li et al., *Topic modeling for short texts with auxiliary word embeddings*. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, 2016, pp. 165-174.

[19] C. Li et al., *Enhancing topic modeling for short texts with auxiliary word embeddings*. ACM Transactions on Information Systems (TOIS), 2017, pp. 1-30.

[20] X. Cheng, X. Yan, Y. Lan, and J. Guo, *Btm: Topic modeling over short texts*. IEEE Transactions on Knowledge and Data Engineering, 2014, pp. 2928-2941.

[21] Y. Zuo, J. Zhao, and K. Xu, *Word network topic model: a simple but general solution for short and imbalanced texts*. Knowledge and Information Systems, 2016, pp. 379-398.

[22] X. Quan, C. Kit, Y. Ge, and S. J. Pan, *Short and sparse text topic modeling via self-aggregation*. Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, 2015, pp. 2270-2276.

[23] Y. Zuo et al., *Topic modeling of short texts: A pseudo-document view*. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Fransisco, 2016, pp. 2105-2114.

[24] K. Ondrej and J. Marlin, *Product life cycle in digital factory*. Knowledge management and innovation: a business competitive edge perspective, Cairo, 2010, pp. 1881-1886.

APPENDIX
PRELIMINARY EXPERIMENT

We conduct a preliminary experiment to prove the effectiveness of our method. The dialog data we used is an original data with very limited context and small corpus size. Hence, we would like to validate our method by experimenting it on a widely used dataset. We performed data augmentation processes on Biomedical dataset taken from [11] which has 20 topics, 4498 corpus, 19448 documents, and 7.44 average document length. Then, we perform topic modeling using LDA, BTM, and PTM algorithm. Finally, we evaluate it by calculating their topic coherence value. Fig. 6 shows that data augmentation will improve topic quality on a certain degree.

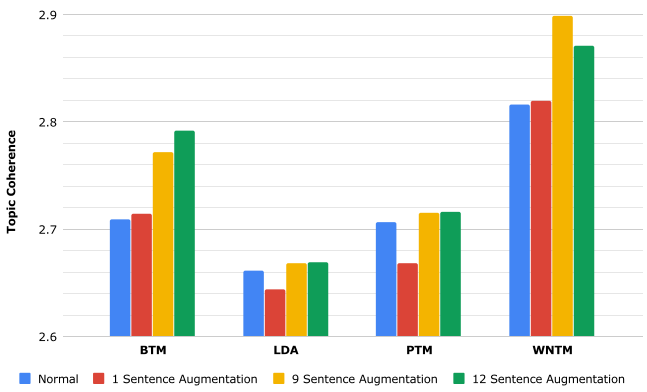


Figure 6. Topic coherence value based on topic model and sentence augmentation process performed on Biomedical Dataset