Analysis on the usage of Topic Model with Background Knowledge inside Specific-themed Discussion Activity

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

***Abstract*—Memek memek memek. This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

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. Furthermore, consensus building is important because it can resolves dispute more effectively by involving people from various levels and departments in an organization [1]. Nowadays, most companies are using consensus building approach on the requirement decision part of their products, hence making such activities as a specific-themed discussion activity.

The practice of consensus building often times still having frequent problems. During discussion activities, various stakeholders who has different personalities and backgrounds are present. This diversity would influence final conclusion [2] and affects their tendency and direction of the discussion ((GOTO, 2019)).

Couple of methods can be implemented inside discussion activity to improve consensus quality e.g. 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 actually recording the whole discussion as a video file or even as simple as taking notes on participant's utterance. Facilitation in a second hand, used to help participants work together by providing an artifact containing the discussion progress which everyone agrees on. Finally, mediation acts to help opposite parties deal with disagreement. One independent participant is needed to resolve disputes with his/her objective point of view.

Few researches has been conducted to improve consensus quality. One initiative takes form by performing recording act on non-verbal aspects of the discussion ((KATAGIRI 2008)). In industrial engineering perspective, another initiative has been proposed as a new framework of short term and intensive workshop facilitation for multi-party stakeholders in Product Lifecycle Management (PLM) strategy planning phase [4]. Both initatives tried to improve the overall discussion activity process while each of it has their own problems. The first initiatives is hard to interpret for general participants because the result takes form as raw Proposal Unit (PUs) while the second one is heavily relied on one external facilitator which might produce biased judgment.

II. Research Problem

In this paper, we tried to resolve the disadvantages found in previous researches. We tried to propose a method that is easy to interpret for general participants and external facilitator. We also tried to produce an objective result to support external facilitator's feedback. Basically, we conducted a digitized approach by analysing dialog data from discussion sessions and analyze it using topic model and background knowledge. We compile the result and validate it on a Japanese company and one external facilitator. The feedbacks we retrieve will be a valuable asset to continue develop this approach. However, a preliminary study regarding this matter has been conducted ((GOTO 2019)) and this research act as the extension of it with approval from the original author.

III. Proposed Method

In this research, we performed digitized approach of dialog data from discussion activity sessions using data augmentation, topic model with background knowledge, and distribution similarity. First, the data will be prepared by simple preprocess and data augmentation. The clean and augmented data will then be experimented by various topic models and hyperparamaters, we picked the best configuration and incorporate it into background-knowledge-backed topic model to generate topic distributions. Finally, we will calculate the distribution similarity by multiple factors and implement it to actual use.

A. Data Augmentation

We took a real life dialog data from discussion sessions which happened for 1-2 hour long. Based on the dataset

characteristics in ((TABLE)), the dataset we used is very poor. Thus, we are using data augmentation techniques to improve dataset quality. We expand the Easy Data Augmentation ((WEI 2019)) by adding additional processes e.g. 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 other sentence s' who has hypernym/hyponym relation with some words inside it.

B. Topic Model with Background Knowledge

We tried to mine latent opinion of participants 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 proposal method of topic model is Latent Dirichlet Allocation (LDA) ((BLEI 2003)), which mostly, the current available topic model is proposed based on that. In LDA-based topic model, the learning process consists of generative 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. While in the sampling process, distributions are evaluated by iterating each word and recalculate the distribution using Gibbs Sampling ((GIBBS REFERENCE)). The graphical notation of LDA topic model is shown in ((FIGURE)), while the algorithm is shown in ((FIGURE)).

In our approach, we realize that our dataset has relatively smaller size compared to common topic model researches. Hence, we compiled various topic models with speciality in short text as suggested by ((QIANG 2019)). The whole list of topic models could be seen in ((TABLE)).

After the experiment is done we can 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 backround-knowledge-backed topic model named Source-LDA ((WOOD 2016)) as the most suitable topic model for our case. In Source-LDA, we could provide backround knowledge data and it will help us improve the topic quality by encouraging the topic label automatically.

C. Distribution Similarity

In this step, we aimed to symbolize the topic disribution into a single value that describes the rate of consensus built. In order to do this, we used distribution similarity calculation using Jensen-Shannon Divergence across all distributions ((submitted1.pdf)). And then the final deliverables of this research will be the final topic distribution with its agreement rate.

IV. Experiment

Dialog data from requirement decision discussion sessions of 4 Japanese companies was succesfully retrieved. Data preprocessing and sentence augmentation is done to clean the data. The comparison of dataset characteristics before and after data modification is shown in ((FIGURE)). Furthermore, the property for each sentence in dataset presentend in ((TABLE)).

Following the data preprocessing step, topic model experiment is conducted on all topic models in ((FIGURE)) with additional experiment on LDA. The hyperparameter and topic evaluation used is based on ((QIANG 2019)) which we used topic coherence since our dataset is raw and no labelling has been done for the dataset. The result of topic model experiment is shown in ((FIGURE)), while the actual effect of sentence augmentation process towards topic coherence value is shown in ((FIGURE)).

From the result, we can conclude that xx sentence augmentation processes gives the best and most consistent result compared to others. Finally, we picked xx topic model with xx hyperparameter and xx sentence augmentation processes as the best configuration.

The next step is to incorporate this configuration into Source-LDA. The utilized dataset was taken from discussion sessions which are part of Product Lifecycle Management (PLM) practices ((PLM REFERENCE)). Hence, PLM topics is used as the background knowledge data. The best choice for this is PTC Value Roadmap ((PTC REFERENCE)) because it contains a whole 26 PLM Topics with the definitions of each. THe characteristics of this dataset is shown in ((TABLE)). ((FIGURE)) shows the value of topic coherence relative to the number of sentence augmentation process applied to background knowledge dataset.

As the final configuration, xx sentence augmentation processes for background knowledge is applied. Since topic distribution has finally obtained. JS Divergence is used to measure the similiarity between topic distributions. The similarity of each discussion sessions can be seen at ((TABLE)).

V. Case Study

VI. Results and Discussion

In this section, the actual result and qualitative evaluation will be presented. ((TABLE)) dst shows the average topic distribution values clustered by various supporting informations.

After we conducted another discussion session with the same set of participants and external facilitator. We retrieved feedbacks as shown in ((FIGURE)). Overall, the participants and external facilitator feels that the usage of topic model with background knowledge in specific-themed discussion activy environment give moderate impact xxx xxx xxx.

VII. Conclusion and Future Works

VIII. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an

independent document. Please do not revise any of the current designations.

IX. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections IX-A–IX-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.

- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

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] 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.
[3] K. Elissa, “Title of paper if known,” unpublished.
[4] 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 IFIP International Conference on Product Lifecycle Management*, Moscow, 2019, pp. 82-93.
[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.