A Publicly Available Indonesian Corpora for Automatic Abstractive and Extractive Chat Summarization

Fajri Koto
Advanced Research Lab
Samsung R&D Institute Indonesia, INDONESIA
fajri.fajri@samsung.com

Abstract

In this paper we report our effort to construct the first ever Indonesian corpora for chat summarization. Specifically, we utilized documents of multi-participant chat from a well known online instant messaging application, WhatsApp. We construct the gold standard by asking three native speakers to manually summarize 300 chat sections (152 of them contain images). As result, three reference summaries in extractive and either abstractive form are produced for each chat sections. The corpus is still in its early stage of investigation, yielding exciting possibilities of future works.

Keywords: chat, summarization, corpora, indonesian

1. Introduction

The growth of Internet around the world engenders new trends for people to actively communicate online by using various application and services involving text, audio, picture, and video. According to *Statista*[1] the active user of WhatsApp[2] - the most well-known mobile chat application in the world, has reached 600 million by 2014. Similar data is also shown by other applications, such as Viber[3], WeChat[4], LINE[5], and KakaoTalk[6]. Consequently, summarization system becomes important to ease people in extracting the information automatically and quickly from chat conversation.

Even though this situation will trigger many researchers to investigate methods and technique related to chat summarization, literatures show that only few studies specifically discuss about it. Whereas, chat summarization has many potential uses. It is applicable to various domains including multinational companies (Handel and Herbsleb, 2002), open source meetings (Shihab et al., 2009) (Zhou and Hovy, 2005), distance learning (Osman and Herring, 2007), and even military (Uthus and Aha, 2011).

Challenge that is mostly discussed is finding suitable chat corpus that can be used for testing and evaluating summarization applications (Uthus and Aha, 2011). Most chat corpora do not have any summaries associated with them to use for a gold standard, making evaluations difficult. To address this issue, we start our investigation by first constructing a corpus containing chat conversation and its associated summary in two forms: abstractive and extractive summary. Extractive summary is obtained by selecting important sentences from the given chat conversation. Whereas, abstractive summarization is a more advanced summarization which involves natural language technique in building new sentences by considering pattern and grammar of language.

In order to achieve research development on chat summarization, we initialize this work by constructing a suitable corpora that can be used for automatic qualitative evaluation. Here we use Indonesian language (Bahasa Indonesia), by considering two factors: 1) Indonesia is the 5th biggest population and the 13th biggest Internet user in the world. According to the International Telecommunication Union (Geneva)[7], Indonesia has more than 35 million Internet users by 2014. Consequently, understanding and providing summary for chat in Bahasa Indonesia is not only useful for user but also for business need like market analysis, politics, and government. 2) Previous researches show only few studies discussing chat summarization, caused by the limited corpora. Therefore we construct the first ever Indonesian corpora for chat summarization by employing three native speakers to manually build the summary.

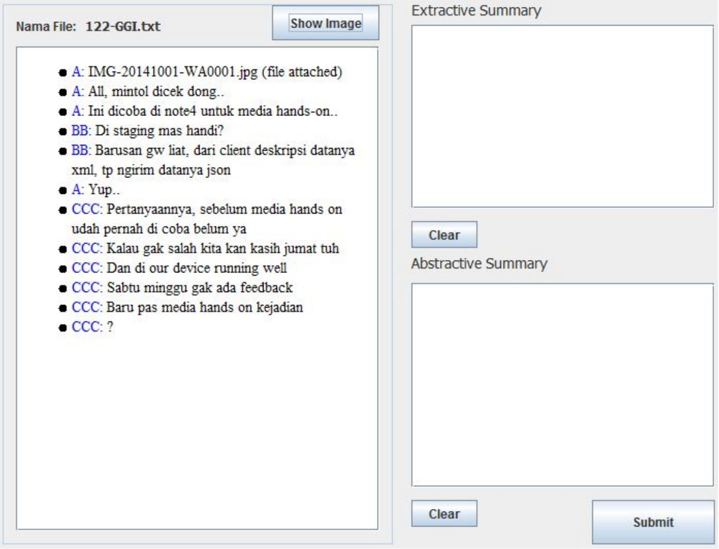
The rest of this paper is structured as follows. Section 2 summarizes some related works of chat summarization. Section 3 provides the construction of our dataset. The analysis of dataset is also discussed in Section 4. Finally conclusion are drawn in Section 5.

2. Related Works

Research related to text summarization has been done in different text genres such as news articles (Lee et al., 2005), scientific articles (Teufel and Moens, 2002), and blogs (Hu et al., 2007). Edmundson (1969) and Luhn (1958) have used features such as average term frequency, title words and cue phrase for scoring sentence to create summary by applying supervised classification. Marcu (1998) used the structural aspect of the text to discover the relationship between the segments of the text.

In case of advanced summarization, the work of abstractive summarization is still limited because of its difficulty to construct new sentences by paraphrasing. Ganesan et al. (2010) performed graph-based summarization framework (Opinosis) that generates concise abstractive summaries of

[1] http://www.statista.com/
[2] https://web.whatsapp.com/
[3] http://www.viber.com/
[4] http://www.wechat.com/
[5] http://line.me/en/
[6] http://www.kakao.com/talk
[7] http://www.itu.int/



Nama File: 122-GGI.txt

- A: IMG-20141001-WA0001.jpg (file attached)
- A: All, mintol dicek dong..
- A: Ini dicoba di note4 untuk media hands-on..
- BB: Di staging mas handi?
- BB: Barusan gw liat, dari client deskripsi datanya xml, tp ngirim datanya json
- A: Yup..
- CCC: Pertanyaannya, sebelum media hands on udah pernah di coba belum ya
- CCC: Kalau gak salah kita kan kasih jumat tuh
- CCC: Dan di our device running well
- CCC: Sabtu minggu gak ada feedback
- CCC: Baru pas media hands on kejadian
- CCC: ?

Figure 1: User Interface of Annotation Program

highly redundant opinions. Genest and Lapalme (2012) developed and implemented a fully abstractive summarization methodology in the context of guided summarization. Although fully abstractive summarization is a daunting challenge, their work shows the feasibility and usefulness of this new direction for summarization research.

Despite the good amount of research in text summarization, works on chat summarization is still very limited. Similar study that resembles chat summarization is conversation of meeting (Hsueh and Moore, 2009) and email (Muresan et al., 2001). However they are different with chat summarization that does not have sound and transcription. Chat conversation tends to be shorter, more unstructured, containing more spelling mistakes, hyperlinks, acronyms, that makes traditional Natural Language Processing (NLP) difficult to apply (Uthus and Aha, 2011).

The most relevant work has been done by Zhou and Hovy (2005) that investigated summarizing chat logs in order to create summaries comparable to the human made GNUe Traffic digest. The corpus used is Internet Relay Chat (IRC) that discusses GNUe, one of the most famous free/open source software projects. The GNUe archive is claimed suitable for chat summarization purpose because each IRC chat log has a companion summary digest written by project participants as part of their contribution to the community. However, the log contains complex structure of the dialog since it talks about question and answering of technical problem like Stack-Overflow[8]. In other words, it's conversation is not as natural as our chat corpus that mostly talks about daily conversation.

A more advanced discussion about chat corpus was discussed by Uthus and Aha (2013). In their paper, they present existing Ubuntu chat corpus as a big data source for multiparticipant chat analysis. However there is still challenge to apply chat summarization, caused by the availability of gold standard for evaluation.

Therefore, to accelerate the development of chat summarization, we make our corpus publicly available. We use some group conversations, and then divide the chat document accordingly based on 2 hours of conversation. The gold standards was built by employing three native speakers to manually build the abstractive and extractive summary. It is also worth to consider that some of our chat documents contain pictures that can be possibly used to boost the summarization by performing image understanding.

3. Data Construction

As discussed in previous section, we utilize WhatsApp, one of the most well-known online instant messaging application, to construct the summarization corpora. WhatsApp is a mobile application that enable user to send/accept text, audio, picture and video to/from another user. WhatsApps also provides a group feature that enable user to conduct multi-participant conversation in a forum.

We argue that logs of multi-participant chat are more suitable than one-on-one conversation in term of data for summarization. Vary and active conversations are more often found in multi-participant than one-on-one conversation. Therefore, we utilized 10 logs of chat conversation group as our raw dataset. It was obtained by utilizing ”Email chat” feature on WhatsApp that enable us to send chat logs to specific email. Most of these logs contain conversations discussing daily activity such as: soccer group, running hobby, faculty organization, family group, and software team development as shown in Table 1.

For each chat logs, we first divided the chat document into sections containing 2 hours of conversations. As result, we obtained 2,281 sections. For data annotation we bounded the data by number of line that equals or greater than 10 in order to avoid sections with few lines. Fewer lines can affect a document having non-extractable summary. In total there are 880 sections, but we only used 300 sections by considering time and cost. 152 of them are sections containing images. Whereas, the rest are obtained by randomly selecting 148 sections from 880 sections.

To ease our freelancers to conduct summarization, we build a simple program as described in Figure 1 with three main

[8] http://stackoverflow.com/

Table 1: Description of our raw data

Logs name	About	#section	Section with #line ≥ 10	
			count	have image
BPH-F	Group of students in Student Executive Council	392	146	6
F-United	Soccer group of college students	344	154	5
P-K	Group of students in Student Executive Council	70	35	2
SR-Runner	Group of runner	63	19	6
CM	Software project development group	284	121	7
Dwi-Fam	Family group	296	37	9
GGI	Software project development group	272	117	38
S-Champion	Software project development group	285	105	12
WAFC	Software project development group	174	82	22
Black-Berry	Software project development group	101	64	45
Total		**2281**	**880**	**152**

functions: 1) to write the extractive and abstractive summary in the given text box; 2) to show image by the button of show image, and; 3) to save the works. Our three freelancers (2 men and 1 woman) are university student and native speaker of Bahasa Indonesia. The same instruction was given before they start doing the works but there is no specific criteria given to construct the summary. As additional information, the name account in our datasets are replaced with unknown characters in order to preserve the privacy of the chat logs owner. One example of extractive and abstractive summary resulted from this stage is given in Table 2.

4. Data Analysis

4.1. Evaluation metric

To apply the automatic evaluation on text summarization, Lin (2004) has proposed ROUGE-N and ROUGE-LCS as the metric to measure and compare how good an approach is. Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summary. Whereas, Longest Common Subsequence (LCS) does not require consecutive matches but in-sequence matches that reflect sentence level order as n-grams. The formula of ROUGE-N is described in Eq. 1 and ROUGE-LCS in Eq. 4.

$$Rouge_N = \frac{\sum\limits_{s\epsilon\{Reff\}} \sum\limits_{gram_n\epsilon S} Count_{match}(gram_n)}{\sum\limits_{s\epsilon\{Reff\}} \sum\limits_{gram_n\epsilon S} Count(gram_n)} \tag{1}$$

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{2}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{3}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \tag{4}$$

Table 2: Example of a section and its associated summary

Chat conversation of F-United:
BB: bro bro. terima kasih buat latian tadi ya bro! (*guys, thanks for the training today!*) A: Sama sama dek makasih juga ya dek ?? (*You are welcome, and thank you too Dek*) CCC: Terima kasih semuanyaa (*Thanks all*) CCC: Ditunggu latian full teamnya (*next game should be a complete team*) A: Siap latihan mingguan jangan jadi arab cuman teori a.k.a wacana ??? (*Don?t be like our mate, Arab in every weekly training, never join???*) BB: iyaa tadi katanya mau tiap kamis tuh latian.. gmn gmn? (*yes, how is our weekly training changed to every thursday?*) CCC: Gw mah berangkat kalo ada amplops (*I will join if there is an envelope*)

Extractive	**Abstractive**
guys, thanks for the training today! *next game should be a complete team*	F-United sehabis melakukan latihan bersama (*F-United just trained soccer together*)

Even though ROUGE is a very famous approach for summarization evaluation, however it is only an n-gram matching of two documents. Consequently, comparing the summary and abstractive reference summary can give a low score. It causes judging the result becomes difficult. Koto et al. (2014) has discussed this issue when working on TED Speech summarization. They proposed Semantic Similarity Checking (SSC) to measure the similarity between two unstructured documents. Below you can find the formula:

$$SSC(D_1, D_2) = \frac{\sum_{i=1}^{i=|D_1|} Sim_{sem}(s_i, D_2)}{|D_1|} \tag{5}$$

In Eq. 5, the D_1 and D_2 represent document of resulting summary and document of reference summary consecutively. The equation simply calculates the average of all semantic similarity score between every sentence s_i in D_1 and D_2. The similarity score $Sim_{sem}(s_i, D_2)$ is also calculated by averaging the semantic similarity score between s_i and all sentences in D_2.

4.2. Original document vs the reference summary

In order to see the agreement score of our freelancers, we compare the reference summary with its original document and provide the result in Table 3. Here we use some parameters comprising the ratio of line number, word number, and character number. We also performed ROUGE calculation between the original document and their reference summary. However, we did not implement SSC to our data, caused by the limitation of Indonesian *Wordnet* to calculate the semantic similarity of words.

Table 3: The comparison between the reference summary and original chat document

Extractive Summary / Original Document				
Parameter	ref-1	ref-2	ref-3	diff
#line	0.229	0.188	0.301	0.075
#word	0.299	0.265	0.387	0.081
#char	0.326	0.305	0.424	0.079
ROUGE-1	0.355	0.313	0.458	0.097
ROUGE-2	0.318	0.282	0.418	0.091
ROUGE-3	0.283	0.251	0.381	0.087
ROUGE-4	0.249	0.219	0.344	0.083
ROUGE-LCS	0.499	0.454	0.581	0.085

Abstractive Summary / Original Document				
Parameter	ref-1	ref-2	ref-3	diff
#word	0.361	0.309	0.141	0.147
#char	0.492	0.449	0.206	0.191
ROUGE-1	0.142	0.092	0.036	0.071
ROUGE-2	0.033	0.022	0.008	0.017
ROUGE-3	0.001	0.008	0.003	0.005
ROUGE-4	0.003	0.003	0.001	0.001
ROUGE-LCS	0.129	0.095	0.048	0.054

In Table 3, column of *ref-1*, *ref-2*, and *ref-3* represent reference summary created by our three freelancers consecutively. The parameter of *#line, #word* and *#char* indicate the ratio number of line/word/char between the summary and original document. We exclude *#line* parameter in *abstractive* table due to its single line. The ROUGE-N and ROUGE-LCS score were also calculated in order to look their similarity. Then the analysis is provided by giving *diff* column that represents the average score of difference between three scores in a row. According to the table, the average of all *diff* scores are 0.0847 and 0.0693 for extractive and abstractive. The scores show that our three freelancer have the similar tendency in building the reference summary.

5. Conclusion

In this work, we have built an Indonesian corpus for chat summarization comprising 300 chat sections. As result of our manual effort, each chat sections in our corpus has had its three kinds extractive and abstractive reference summary that can be used as the gold standard for automatic evaluation. The corpus is free as long as it is used for the development of science and technology. If you want to use this corpus, please contact `fajri.phd@gmail.com` by mentioning your identity and purpose.

6. Acknowledgements

A special thank to Mr. Omar and Mr. Yanuar from Samsung R&D Institute Indonesia who have provided their chat logs for corpus construction.

7. Bibliographical References

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*, pages 340–348. Association for Computational Linguistics.

Genest, P.-E. and Lapalme, G. (2012). Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 354–358. Association for Computational Linguistics.

Handel, M. and Herbsleb, J. D. (2002). What is chat doing in the workplace? In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 1–10. ACM.

Hsueh, P.-Y. and Moore, J. D. (2009). Improving meeting summarization by focusing on user needs: a task-oriented evaluation. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 17–26. ACM.

Hu, M., Sun, A., and Lim, E.-P. (2007). Comments-oriented blog summarization by sentence extraction. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 901–904. ACM.

Koto, F., Sakti, S., Neubig, G., Toda, T., Adriani, M., and Nakamura, S. (2014). The use of semantic and acoustic features for open-domain ted talk summarization. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE.

Lee, C.-S., Jian, Z.-W., and Huang, L.-K. (2005). A fuzzy ontology and its application to news summarization. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 35(5):859–880.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Marcu, D. (1998). *The rhetorical parsing, summarization, and generation of natural language texts. Toronto, University of Toronto*. Ph.D. thesis, Tesis doctoral.

Muresan, S., Tzoukermann, E., and Klavans, J. L. (2001). Combining linguistic and machine learning techniques for email summarization. In *Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7*, page 19. Association for Computational Linguistics.

Osman, G. and Herring, S. C. (2007). Interaction, facilitation, and deep learning in cross-cultural chat: A case study. *The Internet and Higher Education*, 10(2):125–141.

Shihab, E., Jiang, Z. M., and Hassan, A. E. (2009). Studying the use of developer irc meetings in open source projects. In *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*, pages 147–156. IEEE.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Uthus, D. C. and Aha, D. W. (2011). Plans toward automated chat summarization. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pages 1–7. Association for Computational Linguistics.

Uthus, D. C. and Aha, D. W. (2013). The ubuntu chat corpus for multiparticipant chat analysis. In *AAAI Spring Symposium: Analyzing Microtext*.

Zhou, L. and Hovy, E. (2005). Digesting virtual geek culture: The summarization of technical internet relay chats. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 298–305. Association for Computational Linguistics.