Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 229 – 236

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016
9-12 May 2016, Yogyakarta, Indonesia

Topic Summarization of Microblog Document in Bahasa Indonesia using the Phrase Reinforcement Algorithm

Meganingrum Arista Jiwanggi*, Mirna Adriani

Faculty of Computer Science, Universitas Indonesia, Depok, West Java, Indonesia

Abstract

Microblog topic summarization is a part of the challenges to automatically find a topic of any group of microblog posts. This study focused on summarizing Twitter data in Bahasa Indonesia. The main algorithm used in this research is The Phrase Reinforcement Algorithm. This algorithm summarized a group of tweets discussing similar topics using a semi-abstractive approach. As a result of some initial experiments during this study, there are some variations applied in order to obtain summary with a better quality. The evaluation is conducted using human assessment and more than 60% agreed that the summaries have the good grammatical, readability, and informative quality

Keywords: microblog; twitter; topic; summarization

1. Introduction

Users posted in their microblog at the same time with various reasons. In comparison to a news article, the distribution of information in a group of microblog posts with the same topic can vary widely. For instance, a group of posts about a music band's new album could possibly discuss about the album itself, the launching event or even the profile of the band members[1].

Twitter, as one example of a microblog, has a feature called Trending Topic that shows the list of currently popular topics. Twitter displays trending topic in a form of words or phrases. In order to understand the context of a

* Corresponding author. Tel. +62 21 786 3419 ext 3101.
E-mail address: meganingrum@cs.ui.ac.id

doi:10.1016/j.procs.2016.04.054

particular topic, users need to click the topic word and read the related tweets. Rather than showing only a word/phrase, an additional description about what people mainly tweets or discuss would be very helpful for the users.

Given a cluster of microblog posts about a topic taken at a certain timeframe, this study observed the method to automatically create a short description to illustrate the topic. The description is actually the summary of the whole posts in that group. Therefore, this research was basically a study about microblog topic summarization. There are two kinds of summarization based on the approach to compose a summary which are *abstractive*—by composing one new sentence based on all original sentences— and *extractive*—by choosing the best sentence representing all sentences[2].

We conducted a set of experiments about topic summarization using Twitter data in Bahasa Indonesia. Bahasa Indonesia is the main language spoken in Indonesia with more than 250 million speakers. Despite of the huge number of speakers, the data in Bahasa Indonesia is categorized as under-resourced since the study on Bahasa Indonesia are still on early stage. The summary is constructed using semi-abstractive approach and The Phrase Reinforcement Algorithm (TPRA)[4]. The approach was considered a semi-abstractive since it created a new sentence by combining all tweets but it could not be classified as a fully abstractive approach as the process of arranging the result did not utilize the structure of a sentence based on the Natural Language Processing (NLP).

The main contribution in this study we intended to improve the quality of the summary results. The initial experiment was conducted using a similar approach as the earlier work[4]. The result of the initial experiment was that we need to improve some particular aspects. We proposed some methods to improve those aspects such as applying normalization, additional flags in the graph, and also a method to handle different topics in one result sentence. We did some experiments using the variation of the proposed methods. In this research, the evaluation was conducted using human assessment to assess the quality of the summary through some factors such as readability, grammaticality and meaning[3].

2. Related Works

2.1. Hybrid TF-IDF

Scoring is a mandatory step for summarization process using extractive approach. The score will be used to compare the candidates of the best summary. The Term Frequency and Inverse Document Frequency (TF-IDF) is a classic method that is still popular until now. Usually, the researchers modify the method to fit the data.

How often a word occurs in the document shows the degree of importance of that particular word in a text document. The words having less occurrences in a text document can be scored lower than those that have more[5]. TF is the number of occurrence of a particular word in a document. IDF can be calculated as follows:

$$idf_i = \log(\frac{D}{d_i}) \quad (1)$$

where D is the total documents and d_i is the number of document containing the words i.

In this case, we would like to use TF-IDF for Twitter topic summarization. We firstly need to make a clear definition about which part of the Twitter data can be considered as one document. If a document is defined as all tweets in the data, then the calculation of TF will run normally whereas the IDF calculation will face a problem since there is only one document. Meanwhile, if one document is defined as one tweet, then there will be a problem while calculating the TF since most of the words has a very small TF due to relatively short length of one tweet. Considering both reasons, in our case, we need to use a new method called Hybrid TF-IDF[4].

The Hybrid TF-IDF formula is written as follows:

$$\boxed{\begin{aligned} W(S) &= \frac{\sum_{i=0}^{\#WordsInSen\ tence} W(w_i)}{nf(S)} \\ W(w_i) &= tf(w) * \log_2(idf(w_i)) \\ tf(w_i) &= \frac{\#Occurences\ \ OfWordInAl\ \ lPosts}{\#WordsInAll\ \ Posts} \\ idf(w_i) &= \frac{\#SentencesI\ \ nAllPosts}{\#SentencesI\ \ nWhichWord\ \ Occurs} \\ nf(S) &= \max[MinimumThr\ \ eshold\ , \#WordsInSen\ \ tence\ ] \end{aligned}} \tag{2}$$

Note :

- W(S) : the weight of a sentence
- '#' : the number of
- nf : the normalization factor
- w_i : the i-th word.

The formula shows how we define one document -- that is called a hybrid document -- as we calculate TF is all tweets, but on IDF case, one document means one tweet In the previous research, Hybrid TF-IDF has been used to construct a summary from a set of tweets using the extractive approach[6].

2.2. The Phrase Reinforcement Algorithm

The Phrase Reinforcement Algorithm[4] constructs a summary using semi-abstractive approach. This algorithm utilizes a graph to model the input data. It creates a new sentence by combining the sub-graphs with the highest number of occurrence in the document. It cannot be classified as a fully abstractive approach as the process of arranging the result does not utilize the structure of a sentence based on the Natural Language Processing (NLP).

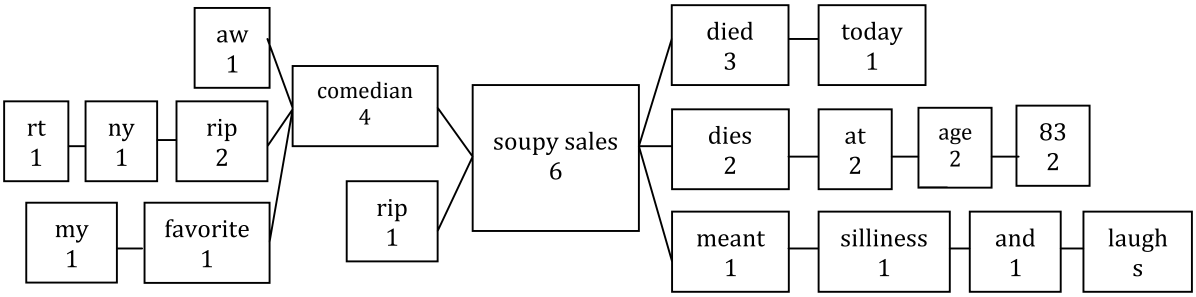


Fig. 1. A graph with 'Soupy Sales' as the root node[4]

The first step is to build a graph of words from the tweets. Fig. 1 shows the example of a graph formed from a group of tweets about "Soupy Sales" (a comedian name)[4]. The root of the graph is a topic word/phrase. As an example showed in Fig. 1, a phrase "Soupy Sales" will be the graph root. For each node *k*, the words occurring before *k* will be arranged at the left side respectively, while the words occurring after *k* will be arranged at the right side. The number of occurrence for each word will be placed in the node. The words that occur less than twice will be cut from the graph.

The score for each node is calculated as follows:

$$\boxed{Weight(Node) = Count(Node) - Dis\tan ce(Node)*^{b}\log Count(Node)} \tag{3}$$

Count(Node) is the number of occurrence of the word in a node. *Distance(Node)* is the distance between the root and the corresponding Node, e.g. the distance of the Node "comedian" with the root "Soupy Sales" is 1, while for the

Node “today” is 2, and etc. Note that b could vary depending on the length of the desired summary result. We can assign a small value for b (e.g. 2 or natural number *e*) to create a short summary with higher precision, while a bigger value (e.g: 100) for a long summary with higher recall. The stopwords should not be included in the scoring.

The candidate for the best sub graph will be explored using the Depth First Search (DFS) method. The DFS starts from a certain sub graph until its maximum depth before moving to another sub graph. After finding the candidates, the final summary will be arranged based on the total weight in both the left and right sub graphs.

3. Experiments

3.1. Data

The tweets were streamed based on a list of queries. To make sure that the data would only be in Bahasa Indonesia, the query words were also in Bahasa Indonesia. Some queries containing English words were still used, such as “persibday” (a national soccer club) and “thankyoustellajkt48” (a national singer group), since both topics specifically talked about local issues in Indonesia.

The data were retrieved at 4 Nov – 23 Dec 2013. A streaming duration was ranging from 1 hour to 4.5 hours. From about 50 topics with more than 100.000 tweets, there were only 30 topics selected for experiments. Those topics were selected based on their representativeness to the various categories, such as politic, law, technology, culture, sports, and entertainment. The number of tweets in each topic varies with a range of 50-16000 tweets.

3.2. Data Preprocessing

The preprocessing of the Twitter data was done by removing URL links, removing non-alphanumeric characters, and transforming the format of letters in lowercase. In this step, we also created a list of stopwords in Bahasa Indonesia containing more than 1000 words. The indicators of stopword were as follows:

- Contain less than 3 letters, except “RT”
- English words with less informative meanings, e.g. : follback, with, like, the
- Interjection and its variations, e.g: *hah, hahh, hei, hey, loh*
- Pronoun and its variations, e.g.: *bu, pak, kamu (km,kmu,qm), gue (gw,g)*
- The other words that often occur but do not have any semantic meanings and its variations, e.g.: *haha (hahaha,hahahaha), wk(wkwkwk), yang(yg)*

The stopwords often dominate the content of a tweet. Removing the stopwords could possibly affect the structure of a sentence in the summary. Therefore, the stopwords would be ignored during the scoring process.

3.3. Data Scoring

The Hybrid TF-IDF was used for data scoring. The value of b log factor chosen for this research was 20 to create neither too short nor too long summary. The initial observation showed that there were many tweets containing only repetition of words especially the trending topic words. This kind of tweet was not informative but able to gain a high score. e.g.: “*#persibday #persibday #persibday #persibday #persibday #persibday*”

Therefore, in this research the score of a sentence was calculated as the total of the score of unique words. e.g.: The Hybrid TF-IDF score for the words: “I” = 1, “eat” = 2, and “pizza” = 3. It implied that:

- The total scores for a sentence “I eat pizza” is 1+2+3 = 6.
- The total scores for a sentence “pizza pizza pizza” is 3.

3.4. Summary Construction

The input data were modeled in a graph and processed further using The Phrase Reinforcement Algorithm. The Depth First Search method is utilized to find the best sub-graph. There were several new variations being tested in this study. While evaluating the result of the initial experiment, we found some problems while applying the algorithm to the data. Therefore, these methods below were proposed to improve the quality of the summaries.

3.4.1. Additional Start-End Flags

In the earlier research, a node would be cut from the graph if the word occurrences are less than 2. Based on this rule, there was a possibility of sub-graph with less important information that still appear as it met the threshold.

Let have one summary as an example:

RT @dhonneysinatra: Beda ceban doang ame swota RT @kompascom: **"Airport Tax" di Bandara Halim Rp 30.000**

The italic part shows the comment that is less important and the main information is marked with the bold parts. By searching through the data, there were more tweets containing only this part:

RT @kompascom: "Airport Tax" di Bandara Halim Rp 30.000

These findings gave an idea to modify the graph by adding the start and the end flag as the indication of a start and an end of a tweet. Let again take an example based on the previous one as follows:

- RT @kompascom: Airport Tax di Bandara Halim Rp 30.000 (*repeated 4 times as retweets*)
- RT @dhonneysinatra: Beda ceban doang ame swota RT @kompascom: Airport Tax di Bandara Halim Rp 30.000 (*repeated 3 times as retweets*)

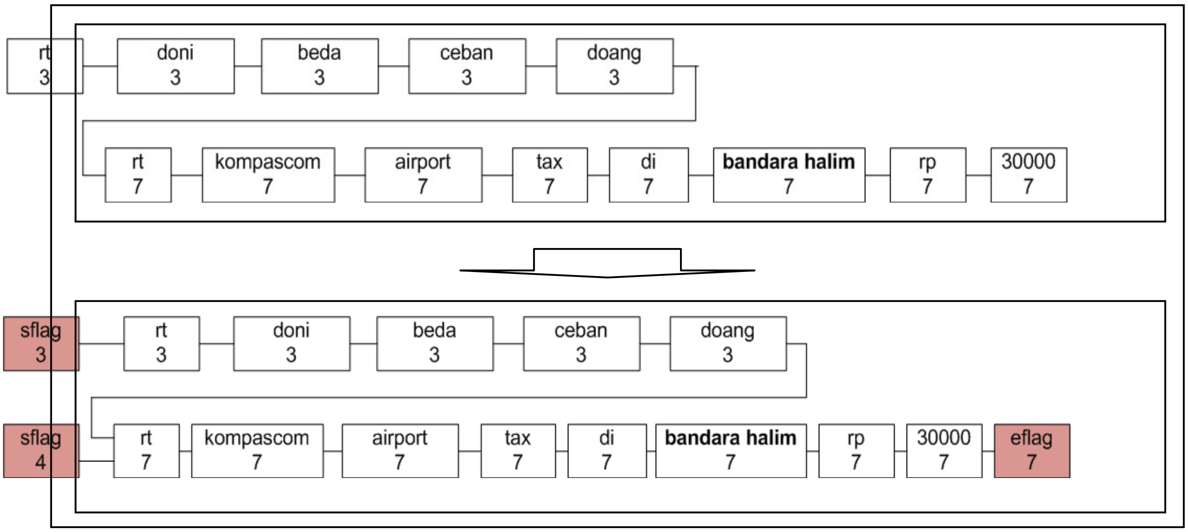


Fig. 2. Example of a graph before and after modification

3.4.2. Negative Scores

Negative scores might occur as a result of the candidate sub-graph that was too long so that the reduction factor in the equation (3) might dominate. One of the effects was the short tweet that might be less informative than the longer one would likely be selected. The formula after normalization became as follows:

$$\boxed{Weight(Node) = Count(Node) - \frac{Dis\tan ce(Node)^{*b} \log Count(Node)}{Length(Sub - Graph)}} \quad (4)$$

3.4.3. Different Topics in the Sub-Graphs

Since the calculation was done separately in the left and right sub-graphs, the summary could contain mismatch topics in the left and right side of the root word. It was likely possible considering there might be more than one sub-topic being discussed about a particular topic.

To resolve this problem, we conducted some additional steps while running the DFS. While selecting the candidates in the left side sub-graphs, we also checked the similarity of the discussion topic in the left and right side sub-graphs. The candidate summary in the right side sub-graph having at least 50% similar parts to those in the left side would be kept. The final summary was arranged based on the candidate pairs with the best total score.

3.5. Data Preprocessing

The aim of postprocessing was to write the summary back into its initial format. The result would help to reduce ambiguity so that users were able to understand the summary contents well.

4. Results

The summary results were evaluated manually using human assessment. The researcher took a part on conducting the initial evaluation. This step produced several candidates to be assessed further by a group of evaluators.

We recruited 37 random evaluators vary from under graduate and graduate students to evaluate the summaries. The evaluators were asked to fill the evaluation form. Basically, they evaluated the result based on the three factors: readability, grammaticality and meaning. The evaluation form was designed to assess how well the summary met those criteria according to each evaluator. The form contained 2 main tasks as follows:

1. The evaluator was asked to determine the topic of the summary by mentioning a topic word/phrase

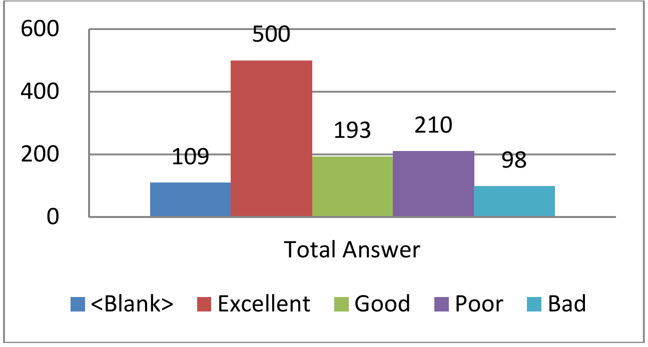
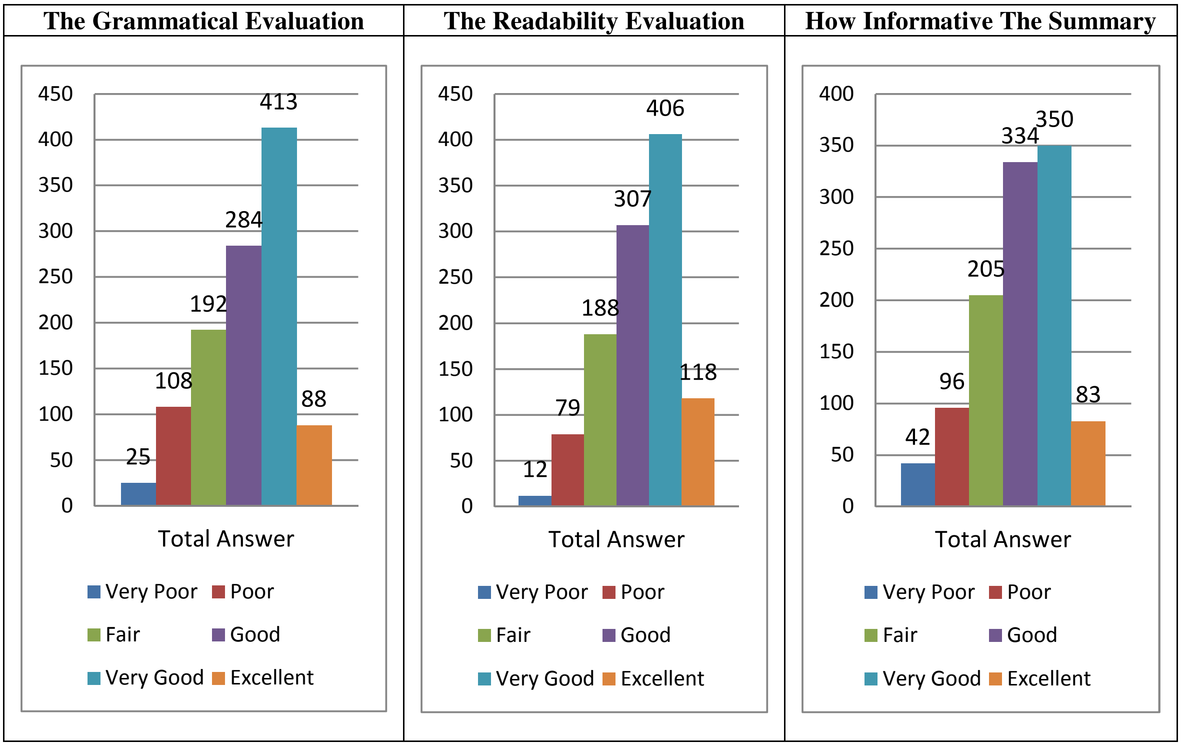


Fig. 3. The category of the total answer for the 1st evaluation task

This evaluation aimed to observe how representative a summary to the real discussion topic. We grouped the evaluation results based on how close the evaluator's written word with the real query word. Assuming that the positive answers were in the "excellent" or "good" group, there are 62.43% of the answers that were positive showing that the summary was good enough to represent a particular topic.

2. The evaluator was asked to give a score of the quality of the summary on 3 indicators: grammatical, readability, informative.

The evaluators need to score the quality of the summary on a 1-6 Likert scale for each indicator. Mostly they scored 5 as the evaluation for grammatical quality of the summary results. The trend showed that the evaluators reacted positively to the grammatical quality of the result with 70.73% of the answers were between 4 and 6 of Likert scale. Mostly they also scored 5 as an evaluation for the readability of the summary results. It indicated how easy the content of the summary to be understood by the readers. The trend showed that the evaluators also reacted positively to the readability of the summary with 74.87% evaluators scored between 4 and 6. Most evaluators also scored 5 as an evaluation for how informative the summary results are. The trend showed that they thought the summary results had given good enough information with 69.1% of the answers were between 4 and 6 of Likert scale.


The Grammatical Evaluation

The Readability Evaluation

How Informative The Summary

Fig. 4. The category of the total answer for the 1st evaluation task

5. Conclusion

We report our study about topic summarization on the microblog written in Bahasa Indonesia. We mainly used The Phrase Reinforcement Algorithm to construct the summary with some modifications on the process as the result of the initial observation using our data. There were three proposed methods that were included in the experiments in order to improve the quality of the summary such as applying normalization, additional flags in the graph, and also a method to handle different topics in one result sentence. The best result of the experiments was evaluated using human assessment from a group of evaluators. The evaluation showed that there were more than 60% total positive answers at the grammatical, readability and informativeness the summary.

The trending topic especially in Bahasa Indonesia contains general topics which could actually be broken down into several distinct sub-topics. For future work, there will be interesting to conduct some experiments to cluster a group of tweets into some smaller groups with more specific sub-topic. Another issue is about the evaluation method. The gold standard usually is hardly found, so that the evaluation needs to be done using human assessment. We plan to explore other methods to evaluate the topic summarization.

References

1. Xu, R. Grishman, A. Meyers and A. Ritter, "A Preliminary Study of Tweet Summarization using Information Extraction," in Proceedings of the Workshop on Language Analysis in Social Media pp. 20-29, Atlanta,Georgia, 2013.
2. A. Olariu, "Clustering to Improve Microblog Stream Summarization," in 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (pp. 220-226), 2012
3. Nichols, J., Mahmud, J., & Drews, C. (2012). Summarizing Sporting Events Using Twitter. IUI'12 (pp. 189-198). Lisbon, Portugal: ACM.
4. Sharifi, B., Hutton, M.-A., & Kalita, J. (2010). Summarizing Microblogs Automatically. The 2010 Annual Conference of the North American Champter of the ACL (pp. 685-688). Los Angeles, California: Association for Computational Linguistics.
5. Grossman, D., & Frieder, O. (2004). Information Retrieval: Algorithms and Heuristics 2nd Edition. AA Dordrecht: Springer.
6. Sharifi, B., Hutton, M.-A., & Kalita, J. (2010). Experiments in Microblog Summarization. IEEE International Conference on Social Computing (pp. 49-56). IEEE