SUMMARIZING INDONESIAN NEWS ARTICLES USING GRAPH CONVOLUTIONAL NETWORK

Garmastewira Garmastewira & Masayu Leylia Khodra
School of Electrical Engineering and Informatics
Institut Teknologi Bandung, Indonesia
ranggamaste@gmail.com; masayu@stei.itb.ac.id

ABSTRACT

Multi-document summarization transforms a set of related documents into a concise summary. Existing Indonesian news article summarization does not take relationships between sentences into account and depends heavily on Indonesian language tools and resources. This study employed Graph Convolutional Network (GCN) which allows for word embedding sequence and sentence relationship graph as input for Indonesian news article summarization. The system in this study comprised four main components: preprocess, graph construction, sentence scoring, and sentence selection components. Sentence scoring component is a neural network that uses Recurrent Neural Network and GCN to produce scores for all sentences. This study used three different representation types for the sentence relationship graph. The sentence selection component then generates a summary with two different techniques: by greedily choosing sentences with the highest scores and by using the Maximum Marginal Relevance (MMR) technique. The evaluation showed that the GCN summarizer with Personalized Discourse Graph, a graph representation system, achieved the best results with an average
ROUGE-2 recall score of 0.370 for a 100-word summary and 0.378 for a 200-word summary. Sentence selection using the greedy technique gave better results for generating a 100-word summary, while the MMR performed better for generating a 200-word summary.

Keywords: Graph Convolutional Network, Personalized Discourse Graph, ROUGE-2, summarization.

INTRODUCTION

Nowadays, people use online media to retrieve the latest news worldwide. Different with information shared in social media that suffers from originality and authenticity (Olanrewaju & Ahmad), the reported news articles often share redundant information because an event will be reported multiple times by different media portals. Multi-document summarization will take these topic-related articles as input and produce a concise summary. An ideal summary contains common information of the news topic and unique information of each article (Goldstein et al., 2000).

There are two common techniques for automatic summarization, which are extractive and abstractive summarization. Extractive summarization extracts important sentences of each document, while abstractive summarization collects important information of documents which are later written in the form of new sentences as the summary (Jurafsky & Martin, 2014). While an ideal human-made summary is an abstractive summary, extractive summarization research often gives a small difference in performance compared to abstractive research that requires more resources and tools (Galanis et al., 2012). Thus, this paper chose to use the extractive approach.

Most extractive summarization research uses supervised learning to determine salient sentences. Kupiec et al. (1995) built one of the earliest extractive summarization systems using Naïve Baye classifier. Each sentence is transformed into simple features such as sentence length and position. Then, a model is trained to classify each sentence as important or the opposite. Hong and Nenkova (2014) created another extractive summarization system based on word importance. Sentence salience is determined by the importance of the sentence’s words, and each word is represented as thousands of features. While this research has one of the best performances in DUC 2004 summarization tasks, it requires many resources and careful feature engineering.
Recent summarization research began to use the powerful artificial neural network to generate summaries. Instead of doing extensive feature engineering, most summarization systems using neural networks only require low-level features as input. Rush et al. (2015) and Zhang et al. (2018) employed sequence-to-sequence architecture to generate single-document abstractive summaries with only word embedding as input. Cao et al. (2017) and Zhang et al. (2017) used Convolutional Neural Network (CNN) for multi-document summarization systems without using any hand-crafted features.

Some multi-document summarization research also utilized sentence relationships represented as graphs. Erkan and Radev (2004) computed sentence importance based on the concept of eigenvector centrality in a cosine similarity sentence graph. Christensen et al. (2013) generated a graph depicting discourse relations between sentences to produce a more coherent summary. Recently, Yasunaga et al. (2017) introduced the usage of Graph Convolutional Network (GCN) (Kipf & Welling, 2016) to their neural network architecture. By using GCN, the system will consider sentence relationships when estimating scores of a sentence. Each work shows that considering sentence relationship graphs will provide better summary results.

There are several multi-document summarization works for Indonesian news articles. Christie and Khodra (2016) proposed an abstractive summarization system employing sentence fusion with word graphs. Reztaputra and Khodra (2017) proposed a summarization system based on the sentence structure of subject, predicate, object, and complement (SPOC). While both do not require training data set, both are very dependent on Indonesian language processing tools and resources. As an example, some bad summaries generated by Reztaputra and Khodra (2017) were often caused by SPOC extraction mistake attributed to the quality of the Indonesian dependency tree used. Although employing sentence relationships have improved the performances of summarization systems (Erkan & Radev, 2004; Christensen et al., 2013; Yasunaga et al., 2017), existing Indonesian multi-document summarization works have not considered relationships between sentences in generating summaries. Since considering sentence relationships is also crucial to identify important sentences in Indonesian texts, we will show that employing sentence relationships improves the performance of Indonesian summarizers.
This paper investigates whether adapting sentence relationships in GCN (Yasunaga et al., 2017) is able to provide better performance for Indonesian news article summarization. As described before, the system generates extractive summaries without employing feature engineering, and only requires word embedding as input. Word embedding can be easily constructed for any language. We will also show that our system provides better results than abstractive summaries from existing systems.

This paper is organized as follows: Section 2 discusses related works; section 3 describes the proposed solution used in this study; section 4 shows the experiment and evaluation; section 5 discusses the results and finally, section 6 presents the conclusion and future plans.

**RELATED WORKS**

There are two main steps in extractive summarization which are sentence scoring and sentence selection. A sentence with a higher score is more likely to be a part of the summary. Kupiec et al. (1995) used Naïve Bayes to classify a sentence represented with surface level features to ascertain whether it was salient or not. Hong and Nenkova (2014) worked on word-level features that were mostly hand-crafted and determined the salience of a sentence as the average importance of each word in the sentence. Sentence selection then selects sentences that had previously been scored. Hong and Nenkova (2014) used a greedy approach by selecting sentences with higher scores first. Other summarization research used selection techniques based on certain heuristics like Integer Linear Programming (ILP) (Galanis et al., 2012) and Maximum Marginal Relevance (MMR) (Goldstein et al., 2000). For example, MMR ensured that selected sentences were relevant to sentence candidates but not redundant to the summary.

The Indonesian news article summarization system created by Christie and Khodra (2016) and Reztaputra and Khodra (2017) were modifications of the clustering-based summarization approach proposed by Sarkar (2009). Figure 1 shows Sarkar’s summarization architecture. Similar sentences are grouped into a cluster. Each cluster is ordered so that the most important cluster will be the first one to be put in summary. From each cluster, a representative sentence is selected to be added to the summary.
Christie and Khodra (2016) did not use representative selection. Instead, they generated all possible sentences by sentence fusion utilizing word graphs that depicted the relation of words in sentences. Reztaputra and Khodra (2017) did not cluster sentences based on its similarity; they instead clustered sentences based on its extracted subject and object similarity. Similar to Christie and Khodra (2016), it generated all possible sentences based on the clusters. Both depend on summary generation steps to select salient sentences. Christie and Khodra (2016) used Integer Linear Programming (ILP) summary generation while Reztaputra and Khodra (2017) used Maximal Marginal Relevance (MMR). However, both approaches are very dependent on Indonesian language resources. For example, Reztaputra and Khodra (2017)'s incorrect generated sentence clusters are mostly caused by the quality of the Indonesian language dependency tree. In addition, clustering approaches often miss important sentences due to incorrect clustering. For instance, two important sentences are clustered into one, but only one sentence out of this cluster is picked out.

Yasunaga et al. (2017) proposed a multi-document summarization using neural networks that can accept sentence relationships graph as additional input using GCN. First, each sentence enters the Recurrent Neural Network (RNN) to produce a sentence embedding. Next, the GCN accepts both sentence embeddings and sentence graph as input to produce better sentence embeddings whereby values

---

**Figure 1.** The framework of clustering-based summarization.
are adjusted to consider sentence relationships. Sentence salience score is then obtained based on its sentence embedding and cluster embedding. Summaries are then composed by choosing high-score sentences greedily.

**PROPOSED METHOD**

This study adapted the summarization method of Yasunaga et al. (2017) to improve the performance of summarization system for Indonesian news articles. Fig. 2 shows the overview of the adapted automatic summarization system. There are four major components, which are preprocessing, graph construction, sentence scoring, and sentence selection. The sentence scoring component applies a neural network architecture that has several sub-components.

![Figure 2. Architecture of proposed summarization system.](image)

**A. Preprocessing**

This component performs sentence segmentation and tokenization of each sentence. We also removed quotations because Reztaputra and Khodra (2017) had shown that doing so improved performance. The results of preprocessing are used as the input for both, the sentence scoring and graph construction components.
B. **Graph Construction**

This component forms a sentence relationship graph for each new topic. In the graph, a node is a sentence and an edge connecting two nodes shows how strong both sentences are related. We used three graph representations used in Yasunaga et al. (2017) as follows:

1. **Cosine graph.** The weight of edge between two sentence nodes is tf-idf cosine similarity of both sentences. Therefore, the maximum weight of an edge is 1.0 if both sentences are exactly the same. Any edge with a weight less than 0.2 are removed, i.e. the weight is changed to 0.0.

2. **Approximate Discourse Graph (ADG)** as proposed by Christensen et al. (2013). Unlike the cosine graph, this graph tries to illustrate the relationship between sentences based on discourse relations. While Christensen et al. (2013) used five aspects; we used only three aspects to determine sentence relationships due to the limitations in resources for the Indonesian language. The three aspects that we used were discourse markers, coreference resolution, and entity continuation.

We used 100 Indonesian discourse markers that depicted various discourse relations, e.g. “walaupun” (although), “begitu pula” (likewise), and “akibatnya” (consequently). For the conference resolution, we used a very simple rule-based approach in which two sentences were connected if the latter sentence contained a pronoun. Lastly, entity continuation relies heavily on Indonesian Named Entity Recognition (NER) proposed by Wibisono and Khodra (2018). Figure 3 shows an example of ADG created for three sentences.

![Figure 3. ADG graph representation example.](image-url)
3. **Personalized Discourse Graph (PDG)** proposed by Yasunaga et al. (2017) is a graphic representation which modifies an ADG edge’s weight by infusing personalization scores of both connected sentences to the edge’s weight. With reference to the regression-based approach to summarization (Ren et al., 2018), the personalization score of a sentence is basically a salience score obtained from surface level features of the sentence. We used the same features described in Yasunaga et al. (2017).

C. Sentence Scoring

We obtained the scores of each sentence with the artificial neural network architecture adapted from Yasunaga et al. (2017). The sentence scoring component contained several sub-components as follows:

1. **Sentence RNN** takes each sentence’s token sequence as an input. Each token will be represented as word embedding. The output of this component is that the last hidden state of RNN will be called as sentence embedding.

2. **GCN** takes sentence embedding from sentence RNN and sentence relation graph as input. The output is final sentence embedding that infuses graph relation into its value.

3. **Document RNN** groups the sequences of sentence embedding of each document. The output is the last hidden state of RNN that will be called as document embedding. Then cluster embedding is calculated by averaging all document embeddings.

4. **Score estimation component** is a regular artificial neural network that will take sentence embedding and cluster embedding as input. A sentence’s score is estimated based on its sentence embedding and cluster embedding.

Every single sub-component has weights that need to be determined in order to obtain best results. To determine the best weights, we trained the network on a training dataset in which the target of the network is the normalized ROUGE-score of all sentences shown in Equation (1) (Yasunaga et al., 2017).

\[
R(s_i) = softmax(ar(s_i))
\]

(1)

Where \(r(s_i)\) is the average of ROUGE-1 and ROUGE-2 recall scores of sentence \(s_i\) and \(\alpha\) is softmax distribution constant. Training was done with
back propagation algorithm. We used cross-entropy loss between the sentence score output and averaged ROUGE score as the loss function.

Each sub-component also has hyperparameters. For example, an RNN needs to determine its hidden state length and maximum length of sequence. Softmax distribution \( \alpha \) shown in Formula (1) is also a hyperparameter. This study determined the values of best hyperparameters by conducting an experiment on the validation results.

D. Sentence Selection

For the sentence selection method, we used two techniques as follows:

1. This study utilized a greedy approach to create the summary of a topic base described by Hong and Nenkova (2014). The component then chose a sentence(s) with the highest score as a part of the current summary as long as the summary had not reached its length limit and the sentence was not deemed as redundant.

2. Maximum Marginal Relevance (MMR) was used to generate the summary. MMR was chosen as an alternative because we often found that the sentence with the highest score was an extremely long sentence that did not fit as a summary. We first selected sentence candidates with the highest scores. Formula (2) was used to generate the summary (Goldstein et al., 2000).

\[
MMR = \text{argmax}_{s_i \in C} [\text{sim}(s_i, C) - \text{max}_{s_j \in S} \text{sim}(s_i, s_j)]
\]

Where \( C \) is a set of available sentence candidates and \( S \) is a set of selected sentences. By doing so, MMR ensures that the selected sentence will represent the candidates’ sentences yet not too similar to the current selected sentences.

**EXPERIMENT AND EVALUATION**

This section is comprised of experiment settings, results, and discussion. Experiment settings will explain the dataset details and how to determine the best models. The results will elaborate the best hyperparameters obtained including details of the validation and test data set evaluation. Finally, the discussion part will analyze the results of the summaries produced by each different system.
A. Experiment Setting

Table 1 shows the data set details used in the experiment. To compare this work’s performance in a fair way to previous Indonesian summarization works, we used the exact same test data set used in Christie and Khodra (2016) and Reztaputra and Khodra (2017). The test data set contained five topics with a total of 87 articles and 1,216 sentences. Each topic had both 100-word and 200-word summaries written by one human annotator.

Table 1

| No. | Aspect                     | Train | Validation | Test  |
|-----|----------------------------|-------|------------|-------|
| 1   | No. of clusters            | 50    | 10         | 5     |
| 2   | No. of documents           | 654   | 137        | 87    |
| 3   | No. of sentences           | 11392 | 2462       | 1216  |
| 4   | No. of 100-word summaries  | 2     | 2          | 1     |
| 5   | No. of 200-word summaries  | 2     | 2          | 1     |

Unlike previous Indonesian multi-document summarization works (Christie & Khodra, 2016; Reztaputra & Khodra, 2017), this approach required conducting supervised learning in order to determine the best weights for the sentence scoring neural networks component. Therefore, we had to use training and validation data sets. We chose to build a separate model for 100-word and 200-word summaries for each summary length, and trained four different models where one model was a model without GCN and the other three were models using GCN, each with a different graph representation type. Each model was trained on 50 clusters of topics and validated on 10 clusters of topics. Unlike the test set, each topic contained 100-word and 200-word summaries that were written by two different annotators.

Validation was performed in order to determine the best hyperparameters for each model. A model with particular hyperparameters was better than other models if the generated validation set summaries had better performance. We measured a system summary’s performance by comparing it to annotators’ summaries using ROUGE-2 recall metric (Lin, 2004).

After the best models were obtained, we evaluated each model with test data set that were used in previous Indonesian summarization studies. ROUGE-2 metric recall was also used as a metric. The performance of the greedy selection and MMR selection were compared for each system.
This study also compared the results with three baseline systems which were clustering-based summarization systems. Baseline I is an adaptation system based on Sarkar (2009) for Indonesian articles, baseline II is a summarizer of Christie and Khodra’s (2016), and baseline III is a summarizer by Reztaputra and Khodra (2017).

In addition, we developed a word embedding from an Indonesian news article collection containing 171,923 articles. This study used Word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2016) techniques to generate word embedding.

B. Results

Table 2 shows the hyperparameters of the best models both for 100-word and 200-word summaries. Both models had several differences in the best hyperparameter values. Interestingly, the best 100-word model only required the first 20 sentences of each article as input to obtain the best results. Figure 4 shows the validation loss of system without GCN and system using GCN for each graph representation for 100-word summary models. From the graph, it is very clear that GCN with ADG and PDG systems had decreased the validation loss compared to systems without GCN. However, GCN with cosine graph system had the worst validation system. GCN with cosine graph system also demonstrated an unstable validation loss.

Table 2

Best model hyperparameters

| No | Parameter                               | Best Value          | 100 words | 200 words |
|----|-----------------------------------------|---------------------|-----------|-----------|
| 1  | Word embedding                          | Word2vec            | FastText  |
| 2  | Batch size                              | 1                   |           |           |
| 3  | RNN hidden state length                 | 300                 |           |           |
| 4  | RNN topology                            | Bidirectional LSTM  | LSTM      |
| 5  | Maximum no. of sentences / documents    | 20                  | 50        |
| 6  | Maximum no. of words / sentences        | 150                 | 50        |
| 7  | Use stemming                            | No                  |           |           |
| 8  | Erase stopwords                         | No                  |           |           |

(continued)
Table 3 shows the results of each system for 100-word and 200-word summaries. GCN with PDG and greedy selection system gave the best results for 100-word summaries while the MMR gave much worse results for all systems compared to the greedy selection system for 100-word summaries. Even though it was shown in Yasunaga et al. (2017) that all GCN systems achieved better results, systems without GCN achieved better results than GCN with ADG and cosine graph systems.

For 200-word summaries, GCN with PDG once again gave the best results, but this time with MMR selection. As compared to the 100-word summary results, the MMR gave a slightly better performance for 200-word summaries than the greedy selection. GCN with ADG and PDG systems outperformed...
Table 3

**Evaluation results for 100-word and 200-word summaries**

| No  | System                  | ROUGE-2 Recall |           |           |
|-----|-------------------------|----------------|-----------|-----------|
|     |                         | 100 word       | 200 word  |           |
|     |                         | Greedy         | MMR       | Greedy    | MMR       |
| 1   | No GCN                  | 0.315          | **0.296** | 0.301     | 0.300     |
| 2   | GCN – Cosine            | 0.246          | 0.215     | 0.263     | 0.263     |
| 3   | GCN – ADG               | 0.217          | 0.266     | 0.357     | 0.348     |
| 4   | GCN – PDG               | **0.370**      | 0.204     | 0.323     | **0.378** |

systems without GCN while GCN with cosine similarity graph resulted in a declining performance. Table 4 then compares the performance of the proposed solution to baselines. The system without GCN already performed slightly better than the baselines and GCN with PDG system showed a very significant improvement from the baselines.

Table 4

**Evaluation results compared to baselines**

| No  | System                          | ROUGE-2 Recall |           |           |
|-----|---------------------------------|----------------|-----------|-----------|
|     |                                 | 100 Words      | 200 Words |           |
| 1   | GCN – PDG                       | 0.370          | 0.378     |           |
| 2   | No GCN                          | 0.315          | 0.301     |           |
| 3   | Baseline I – Sarkar (2009)      | 0.297          | 0.292     |           |
| 4   | Baseline II – Reztaputra & Khodra (2017) | 0.280 | 0.288     |           |
| 5   | Baseline III – Christie & Khodra (2016) | 0.231 | 0.319     |           |

**DISCUSSION**

As shown in Table 4, the GCN – PDG system’s ROUGE-2 recall is much higher compared to the previous baselines. Table 5 shows the best summary with a huge ROUGE-2 recall score of 0.703. It can be seen that the summary is almost identical to the reference summary with some grammatical errors. The decrease in ROUGE score was caused by a few unimportant phrases (“In their press release, Tuesday, 21 June 2016”) and different phrases that actually had the same meaning (“to synergize” and “a synergy between”).
Table 5

The best generated summary

| Generated Summary | Reference Summary |
|-------------------|-------------------|
| Dalam siaran persinya, Selasa, 21 Juni 2016, kemitraan Indosat Ooredoo dengan Fujitsu Indonesia akan diperluas ke berbagai sektor industri, termasuk sektor publik untuk memenuhi kebutuhan pelanggan korporasi di Indonesia, tetapi juga perusahaan-perusahaan Jepang yang beroperasi di Indonesia. Kemitraan antara Indosat Ooredoo hadir dalam wujud sinergi pada perangkat, layanan, dan solusi teknologi telekomunikasi milik Indosat Ooredoo, untuk menghadirkan produk dan layanan baru bagi pelanggan korporasi di Indonesia. Ini yang melatarbelakangi Indosat Ooredoo dan Fujitsu Indonesia menandatangani nota Kesepakahan dalam rangka memantapkan kemitraan dengan para pelanggan bisnis yang telah dibangun dan bekerja sama dalam menghadirkan solusi Smart Mobility dan Internet of Things (IoT). | Indosat Ooredoo bersama Fujitsu Indonesia menandatangani Nota Kesepakahan dalam rangka memantapkan kemitraan dengan para pelanggan bisnis yang telah dibangun dan bekerja sama dalam menghadirkan solusi Smart Mobility dan Internet of Things (IoT). Sebagai awal, kerjasama ini akan berfokus pada sektor otomotif dan transportasi dan akan diperluas ke berbagai sektor industri termasuk sektor publik untuk memenuhi kebutuhan pelanggan korporasi di Indonesia tetapi juga perusahaan-perusahaan Jepang yang beroperasi di Indonesia. Kemitraan ini mensinergikan perangkat, layanan, dan solusi teknologi informasi yang mutakhir dari Fujitsu dengan teknologi telekomunikasi terdepan Indosat Ooredoo untuk menghadirkan produk dan layanan baru yang bermanfaat bagi pelanggan korporasi di Indonesia. |

In their press release, Tuesday, 21 June 2016, Indosat Ooredoo together with Fujitsu Indonesia’s partnership will be expanded into various industrial sectors, including public sector to fulfill the need of Indonesian corporate customers, as well as Japanese companies that operate in Indonesia. In this partnership is introduced as a synergy between Fujitsu’s gadgets, services, and IT solutions and Indosat’s telecommunication technology to create new products and services for corporate customers in Indonesia. This is what initiated Indosat Ooredoo and Fujitsu Indonesia to sign a Memorandum of Understanding in order to strengthen their partnership with current business customers and work together to introduce Smart Mobility and Internet of Things (IoT) solution. As a start, this partnership will focus on automotive and transportation sectors and will be expanded into various industrial sectors, including public sector to fulfill the need of Indonesian companies, as well as Japanese companies that operate in Indonesia. This partnership will synergize gadgets, services, and IT solutions that are sophisticated from Fujitsu with Indosat Ooredoo’s foremost telecommunication technology to create new products and services that are beneficial for corporate customers in Indonesia.
On the other hand, Table 6 shows the worst summary results. Time and place information phrases like “Kamis (2/6/2016) (Thursday (2/6/2016))” and “Tangerang Selatan City” are repeated many times. This redundant information fills up the 100-word length limit and therefore blocks other salient sentences from being written in the summary. While the MMR is better on preventing redundancy, it is more prone to select non-salient sentences.

Table 6

The worst generated summary

| Generated Summary | Reference Summary |
|-------------------|-------------------|
| Dalam siaran persnya, Selasa, 21 Polres Tangerang Selatan menyebutkan, sementara ini belum ada informasi terkait korban jiwa dari peristiwa Gedung Panin di Bintaro Sektor 7, Tangerang Selatan, yang roboh, Kamis (2/6/2016) siang. Robohnya salah satupagi dari gedung kembar Menara PermataBank di kawasan Bintaro, Tangerang Selatan (Tangsel), pada Kamis (2/6/2016) siang diindikasikan akibat kegagalan struktur. Sebuah gedung kosong berlokasi di dekat Bintaro Trade Center, Sektor 7, Bintaro, Kota Tangerang Selatan (Tangsel), roboh pada Kamis (2/6), sekitar pukul 15.00 WIB. | Sebuah gedung pencakar langit di kawasan Bintaro Sektor 7, Pondok Aren, Kota Tangerang Selatan (Tangsel) mendadak ambruk, Kamis (2/6/2016) sore sekitar pukul 15.00 WIB. Gedung ambruk ini memang sudah sejak lama dikosongkan karena memiliki struktur bangunan yang tidak baik dan cenderung miring. Pada saat kejadian, gedung ini sedang dalam proses pembongkaran oleh 15 orang pekerja. Namun, saat pekerja tengah beristirahat, tiba-tiba bagian depan gedung ambruk. Tidak ada korban jiwa dalam kejadian ini, tetapi peristiwa ini membuat penutupan arus menuju Pondok Aren dan membuat pengendara yang biasanya bisa langsung belok ke kiri terpaksa harus naik flyover terlebih dahulu lalu memutar jalan. |
| In his press conference, Tuesday, 21, Tangerang Selatan Police Department said that meanwhile, there hasn’t been any information regarding to deaths caused by Gedung Panin accident at Bintaro Sector 7, Tangerang Selatan, which collapsed on Thursday (2/6/2016) afternoon. The collapse of one part of twin building Menara PermataBank at Bintaro Area, Tangerang Selatan (Tangsel), on Thursday (2/6/2016) afternoon is caused by structure failure. An empty building close to Bintaro Trade Center, Sector 7, Bintaro, Tangerang Selatan (Tangsel) city, collapsed on Thursday (2/6) at around 15.00 WIB. | A skyscraper in Bintaro Area Sector 7, Pondok Aren, Tangerang Selatan (Tangsel) city suddenly collapsed on Thursday (2/6/2016) afternoon around 15.00 WIB. This collapsed building is already empty for a long time because it has a bad building structure and looks tilted. When the collapse occurred, it was on a demolition process involving 15 workers. However, when the workers were resting, the front side of the building suddenly collapsed. There is no death reported, but this accident triggered a road closure to Pondok Aren, thus making drivers who usually can turn right directly to use flyover first and then turn around. |
We analyzed how each type of graph would influence the sentence scoring results. The study used “Gempa Dieng” (Dieng Earthquake) topic as an example. Table 7 shows three most important scores generated by GCN with cosine graph system. The bold text shows overlapping phrases between sentences. It is very clear that the top three sentences are talking about the same information. All three sentences also share the same long phrase, which is “Pusat Vulkanologi dan Mitigasi Bencana Geologi (PVMBG)” (Centre for Volcanology and Geological Hazard Mitigation). This boosts the cosine similarity weight between sentences. This may be the reason why the performance of the cosine graph was even worse as compared to no graph because the selected sentences were often similar and did not add new information.

Table 7

*Gempa Dieng topic’s most important sentences of GCN with Cosine Graph System*

| No | Sentence | Score |
|----|----------|-------|
| 1  | Sementara Kepala Pusat Vulkanologi dan Mitigasi Bencana Geologi (PVMBG) Surono mengatakan bahwa gempa yang terjadi di Dataran Tinggi Dieng pukul 19.00 – 19.25 WIB terekam sebanyak 86 kali dengan amplituda maksimum 10 -- 100 milimeter (mm) dan lama gempa 10 -- 70 detik. | 0.043 |
|    | Meanwhile, the head of Volcanology and Geological Disaster Mitigation Center (VGDMC) Surono said that the earthquake at Dataran Tinggi Dieng at 19.00 – 19.25 WIB has been recorded 86 times with maximum amplitude of 10 – 100 millimeters (mm) and duration of 10 – 70 seconds. |       |
| 2  | Same as above (duplicate) | 0.043 |
|    | Data yang dikeluarkan dari Pusat Vulkanologi dan Mitigasi Bencana Geologi (PVMBG) menyebutkan gempa mulai terasa pukul 19.00-19.25 WIB dan telah terekam gempa sebanyak 86 kali dengan maksimum 10-100 mm dengan lama gempa 10-70 detik | 0.027 |
| 3  | The data that is announced by Volcanology and Geological Disaster Mitigation Center (VGDMC) showed that the earthquake started at 19.00 – 19.25 and has been recorded 86 times with maximum amplitude of 10 – 100 millimeters (mm) and duration of 10 – 70 seconds. |       |
Table 8 shows the same for GCN with PDG system where the bold text indicates fulfilled discourse relation aspects. All of the sentences also share entity “Dataran Tinggi Dieng” (Dieng Plateau) to fulfill the entity continuation aspect. Most of the sentences also have discourse markers as the first word, i.e. “Sementara” (“Meanwhile”). Each of the five sentences gives different information. In this way, GCN with PDG system will be able to obtain richer information. GCN with ADG system showed similar results, but it had a bit of tendency to choose non-salient sentences that fulfilled the discourse aspects.

Table 8

| No | Sentence | Score |
|----|----------|-------|
| 1  | *Especially in Kecamatan Bwang area that borders with dataran tinggi Dieng,* said Adi, 34 years old, resident of Desa Rejosari, Kecamatan Bawang, Kabupaten Batang, when contacted by Tempo, Saturday 20 April 2014. | 0.041 |
| 2  | *If we see from that coordinate, he said, the epicenter is estimated to be at Desa Tanji Gugur, Kecamatan Bawang, Kabupaten Batang, and its location is very close to Dataran Tinggi Dieng.* | 0.021 |
| 3  | *One of the public relations staff of Semarang’s National Search and Rescue Agency, Maulana Afandy, said that there were 380 people who were evacuated to Dukuh Bintoro Mulo, Desa Pranten, Kecamatan Bawang akibat gempa di dataran tinggi Dieng.* | 0.014 |
Similar to Yasunaga et al. (2017), we also analyzed the influence of each type of graph representation by measuring the correlation between the sum of incoming edges’ weight of a sentence node and its salience score. Table 9 shows the average correlation results. It can be seen that PDG has a stable correlation value for each topic while ADG can result in a very high and a very low correlation. As the results in Table 9 mostly show high and positive correlations, we can conclude that a sentence that is interconnected with many sentences will most likely have a high salience score.

Table 9

| No. | Data topic                      | Cosine graph | ADG  | PDG  |
|-----|---------------------------------|--------------|------|------|
| 1   | Gedung Roboh                    | 0.344        | 0.717| 0.458|
|     | *Collapsed building*           |              |      |      |
| 2   | Gempa Dieng                     | 0.307        | 0.591| 0.277|
|     | *Dieng Earthquake*             |              |      |      |
| 3   | Kasus Bunuh Diri                | 0.316        | -0.118| 0.283|
|     | *Suicide case*                  |              |      |      |
| 4   | Kerja Sama Indosat-Fujitsu      | 0.120        | 0.011| 0.243|
|     | *Indosat-Fujitsu Collaboration*|              |      |      |
| 5   | Pesawat Salah Mendarat          | 0.250        | 0.449| 0.383|
|     | *Plane landing gone wrong*      |              |      |      |
|     | Total                           | 0.257        | 0.330| 0.329|

CONCLUSION

The proposed adaptation of summarization using GCN has a good performance with an average ROUGE-2 of 0.370 for 100 words and an average ROUGE-2 of 0.378 for 200 words. These results are better compared to previous Indonesian clustering-based summarization research. The best system is GCN system with PDG. Unlike the adapted approach, GCN system with ADG and cosine graph did not perform better compared to the system without GCN. Sentence selection using greedy technique gave better results for generating 100-word summaries, while the MMR performed better in generating 200-word summaries.

Improvements can be made by including other discourse aspects in Indonesian texts for better ADG and PDG representations such as deverbal-noun reference
and more advanced coreference resolution. Like Alias et al. (2017), sentence compression could be considered to reduce the length of each sentence to remove unnecessary words so that the system can take in more compressed sentences as part of the summary.

ACKNOWLEDGMENT

This work was supported by P3MI Institut Teknologi Bandung Research Grant 2018.

REFERENCES

Alias, S., Mohammad, S. K., Gan, K. H., & Ping, T. T. (2017). MYTextSum: A Malay text summarizer model using a constrained pattern-growth sentence compression technique. In International Conference on Computational Science and Technology (pp. 141–150). Springer, Singapore.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5(1), 135–146. arXiv preprint arXiv:1607.04606.

Cao, Z., Wei, F., Dong, L., Li, S., & Zhou, M. (2015). Ranking with recursive neural networks and its application to multi-document summarization. In AAAI (pp. 2153–2159).

Cao, Z., Li, W., Li, S., & Wei, F. (2017). Improving multi-document summarization via text classification. In Proceedings of the 31st AAAI Conference on Artificial Intelligence. 3053–3059.

Christie, F., & Khodra, M. L. (2016). Multi-document summarization using sentence fusion for Indonesian news article. 2016 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA). IEEE. doi: 10.1109/ICAICTA.2016.7803134.

Christensen, J., Soderland, S., & Etzioni, O. (2013). Towards coherent multi-document summarization. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies (pp. 1163–1173).

Erkan, G., & Radev, D. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457–479.
Galanis, D., Lampouras, G., & Androutsopoulos, I. (2012). Extractive multi-document summarization with integer linear programming and support vector regression. *Proceedings of COLING 2012*, 911–926.

Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization* (pp. 40–48). Association for Computational Linguistics.

Hong, K., & Nenkova, A. (2014). Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 712–721).

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing*. London: Pearson.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. In *Proceedings of International Conference on Learning Representations (ICLR) 2017*. arXiv preprint arXiv:1609.02907.

Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4), pp. 485–525.

Kupiec, J., Pederssen, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 68–73). ACM.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches (WAS 2004)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 3111–3119).

Olanrewaju, A. S. T., & Ahmad, R. (2018). Examining the information dissemination process on social media during the Malaysia 2014 floods using Social Network Analysis (SNA). *Journal of Information and Communication Technology*, 17(1), 141-166.

Ren, P., Chen, Z., Ren, Z., Wei, F., Nie, L., Ma, J., & Rijke, M. D. (2018). Sentence Relations for Extractive Summarization with Deep Neural Networks. *ACM Transactions on Information Systems (TOIS)*, 36(4), 39.

Reztaputra, R., & Khodra, M. L. (2017). Sentence structure-based summarization for Indonesian news article. *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*. doi: 10.1109/ICAICTA.2017.8090983
Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 379–389. arXiv preprint arXiv:1509.00685.

Sarkar, K. (2009). Sentence clustering-based summarization of multiple text documents. TECHNIA–International Journal of Computing Science and Communication Technologies, 2(1), 325–335.

Wibisono, Y., & Khodra, M. L. (2018). Pengenalan entitas bernama otomatis untuk bahasa Indonesia dengan pendekatan pembelajaran mesin (Automatic Named Entity Recognition for Indonesian Language using Machine Learning). Seminar Tahunan Linguistik 2018 | Annual Conference of Linguistics. INA-Rxiv.

Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., & Radev, D. (2017). Graph-based neural multi-document summarization. In Proceedings of the 21st Conference on Computational Natural Language Learning. arXiv preprint arXiv:1706.06681.

Zhang, Y., Er, M. J., Zhao, R., & Pratama, M. (2017). Multiview convolutional neural networks for multidocument extractive summarization. IEEE Transactions on Cybernetics, 47(10), 3230–3242.

Zhang, J., Tan, J., & Wan, X. (2018). Towards a neural network approach to abstractive multi-document summarization. arXiv preprint arXiv:1804.09010.