Data-Driven News Generation for Indonesian Municipal Election

Steffi Indrayani and Masayu Leylia Khodra
School of Electrical Engineering and Informatics Institut Teknologi Bandung
Bandung, Indonesia
steffiindrayani16@gmail.com and masayu@stei.itb.ac.id

Abstract—**In order to fulfill the needs of journalistic automation, we develop automatic news generator that accepts structured data and user query and generates Indonesian news article. This paper employs template-based natural language generation in generating Indonesian municipal elections. Based on evaluation using Indonesian news characteristics as the evaluation metric by 15 linguistic experts and 25 active news readers, the average score of the generated news was 3.292 out of 4.**

Keywords-news generation; Indonesian news; template; structured data

I. INTRODUCTION

Over a particular period of time e.g. municipal election or sporting events, a lot of news articles are needed to be produced from a limited number of journalists. In order to address this problem, NLG (natural language generation) has been developed to fulfill the needs of journalistic automation. The Associated Press, one of the largest financial media in the world, has increased the production of articles 12 times using Wordsmith, a template-based natural language generation platform that converts structured data into publishable articles in seconds [1].

Generating news article automatically is known as automated journalism. The advantage of using automated journalism is not only a small amount of time and resources needed to generate news, but also variations of news articles generated in the same topic or news domain. It also ensures objectivity and accuracy of the news as they are based on data [2].

Some researches on automated journalism have been conducted for several languages, such as English, Finnish, and Russian. The news topics vary from the Russian stock market trends [3] to Finland's city elections [4]. Unfortunately, automated journalism has not been applied for Indonesian news article. Thus, this paper aims to develop an automatic Indonesian news generator for municipal election.

Like Leppänen et al. [4], this paper employs template-based approach by adapting Reiter & Dale's NLG architecture [5] and facts representation. We demonstrate the generation of news article of Indonesian municipal elections based on official election committee data and user query.

This paper is organized in the following manner. Section 2 discusses related work about NLG and automated journalism. Section 3 describes the proposed system. Section 4 tells about evaluation and discussion. Section 5 gives the conclusion and future work.

978-1-5386-4804-9/18/$31.00 @2018 IEEE

II. RELATED WORK

Automated journalism is a field that enables a news article to be generated by computer program without or with a little intervention of a human journalist [2]. Automated journalism has advantage that the news is produced faster, more accurate, and more objective. Before automated journalism, there has been computational journalism that used computer program for only collecting, analysis, and organizing input data. In this paper, automated journalism employs data-to-text NLG, that accepts structured data and query as system input and text as system output.

A. Natural Language Generation

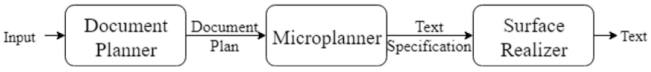

Figure 1. NLG Architecture [5]

Reiter & Dale [5] proposed seven main tasks in NLG, including content determination, document structuring, lexicalization, aggregation, referring expression generation, linguistic realization and structure realization. Those tasks are divided into three modules, i.e. document planner, microplanner and surface realizer. Figure 1 shows the NLG architecture proposed by [5].

Document planner module consists of two tasks, i.e. content determination and document structuring. The first module accepts structured data or knowledge source, communicative goal, and user model. In content determination, the task is to select relevant data or contents based on communicative goal, and summarize or adjust data based on user model or user needs. An example of data summarization is computing the average of daily temperature between 10 AM to 12 PM. The second task determines how contents are ordered or grouped. Output of this module is a document plan that consists of structured relevant data [5].

Sentence 1 : Peter is a good cricket player.

Sentence 2 : Peter is a good chess player.

Result : Peter is a good cricket player **and** Peter is a good chess player.

Figure 2. Aggregation Example [5]

Microplanner module consists of lexicalization, aggregation and REG (referring expression generation) tasks. The first task is mapping data to words or other linguistic representations that express structured data into human language. The second task is to combine several sentence or phrase candidates into single sentence. There are four ways in aggregation, such as simple conjunction, conjunction via shared participants, conjunctive via shared structure, and syntactic embedding. An example of aggregation with simple

conjunction can be seen in Figure 2. The third task is used to reduce the repetition of mentioning the same entity [5].

The surface realizer module consists of linguistic realization and structure realization. In the first task, sentences are generated with the right language morphology. Relevant punctuation, preposition, or auxiliary verbs are added into the sentences. In the second task, sentences are structured into full text [5].

B. *Automated Journalism*

Leppänen et al. [4] built a news generator system that produces Finnish and Swedish news about Finnish municipal elections 2017. They proposed general facts representation for data that are represented in the news. The facts representation consists of entity type, entity, location type, location, value type and value. The type features (entity type, location type and value type) are used to describe the category of the explained features so that the entity, location and value features are not limited to a category or a type only. Leppänen et al. adapted [5]'s NLG architecture as shown in Figure 1. Leppänen et al. also used template-based approach to generate sentence from data structure. According to Leppänen et al., a template is a sentence that has slots to be filled with data and has condition to be fulfilled [4].

Nesterenko [3] developed a system that generates Russian stock news. Similar to Leppänen et al. [4], Nesterenko used template-based approach because of the lack of good stock corpus that could be used for statistical approach. His system also has an analysis module that detects the behavior of stock indexes. The generated news is evaluated using BLEU metric and manual evaluation. The BLEU value appeared to be 0.66, whereas 61% respondents state that the texts are fluent [3].

Besides researches, there are several well-known automated journalism products. Wordsmith [1] is NLG platform that transforms data into narration. Wordsmith uses template-based approach to generate news so that users can tailor the sentence. Associated Press uses Wordsmith to generate quarter financial report to escalate the number of news produced [1].

Los Angeles Times also built a bot called Quakebot that produces earthquake news about three minutes after the real event strikes. Los Angeles Times employed the template-based approach as well [2].

III. PROPOSED SYSTEM

re | https://pilkada2017.kpu.go.id/hasil/2/form_d/t1/25924

SERTIFIKAT REKAPITULASI HASIL DAN RINCIAN PENGHITUNGAN PEROLEHAN SUARA DARI SETIAP KABUPATEN/KOTA I
GUBERNUR DAN WAKIL GUBERNUR TAHUN 2017 PUTARAN KEDUA
KABUPATEN/KOTA : JAKARTA BARAT, PROPINSI : DKI JAKARTA

Terverifikasi

URAIAN		RINCIAN				
DATA PEMILIH DAN PENGGUNA HAK PILIH		CENGKARENG	GROGOL PETAMBURAN	KALIDERES	KEBON JERUK	KEMBANGAN
Data Pemilih						
1. Pemilih terdaftar dalam Daftar Pemilih Tetap (DPT)	LK	177.098	83.034	143.912	120.173	96.211
	PR	172.450	83.858	140.095	120.108	95.668
	JML	349.548	166.892	284.007	240.281	191.879
2. Pemilih terdaftar dalam Daftar Pemilih Pindahan (DPPh)	LK	101	160	86	91	118
	PR	95	146	108	81	100
	JML	196	306	194	172	218
3. Pemilih terdaftar dalam Daftar Pemilih Tetap Tambahan (DPTb)	LK	2.992	802	1.891	1.643	2.095
	PR	3.268	860	2.113	1.924	2.270

Figure 3. Official Sites of Indonesian Elections Committee [6]

The automatic news generator built in this research accepts structured data and query as input. There are five sets of data used in this system, i.e. Indonesian municipal election data, entity fact, single templates, aggregated templates and Indonesian locations data. The example of Indonesian municipal election structured data source taken from Indonesian election committee official site [6] can be seen in Figure 3.

Tahun (Year)	:	2017
Tingkat (Level)	:	*Gubernur* (Governor)
Daerah Pemilihan (Electoral District)	:	DKI Jakarta
Putaran (Cycle)	:	2
Lokasi Pencoblosan (Voting Location)	:	DKI Jakarta
Fokus (Focus)	:	*Pasangan Calon* (Candidate)
Pasangan Calon (Candidate)	:	All
Information yang Dibutuhkan (Needed Information)	:	*Jumlah Suara, Persentase Suara* (Number of candidate votes, Percentage of candidate votes)

Figure 4. User Query Example

Query represents user's needs about what news should be generated. Since our system is expected to produce news articles about Indonesian municipal elections (president, governor, regent or mayor elections), system query consists of year, level, location, and cycle of the elections, needed information and focus of the news. An example of user query can be seen in Figure 4.

The output of our system is news article in Indonesian language. The generated news article only contains information that has been stated in the data. Statement or quotation from other people is not included. Generated news article has one of three defined news focuses (candidate, party or elector). In addition, the news produced has inverted pyramid structure and a paragraph in text has at least two sentences.

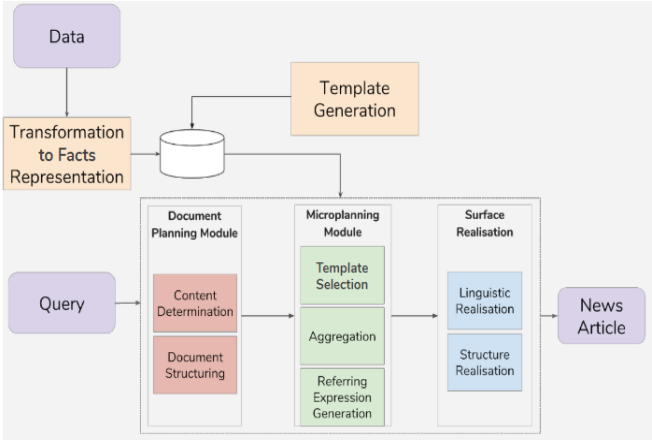

Figure 5. Process flow in generating news article

The system adapts Reiter & Dale [5] NLG architecture and uses template-based approach. Our system also adapts facts representation from Leppänen et al. [4]. In general, the processes involved in generating news can be seen in Figure 5. Next, we will discuss about all processes and modules in the system.

A. Template Generation

As mentioned before, template is a sentence that has slots to be filled and conditions to be fulfilled. There are two types of template, standard or single template and aggregated template.

Single template is a sentence that has only one value type condition to be fulfilled. A single template has to have a value slot and a location or event slot. It also has at least entity type and value type conditions. A template can have another template as a couple that will be used in determining aggregated template. Examples of single template can be seen in Table 1. These templates are also a template couple.

TABLE 1. SINGLE TEMPLATE EXAMPLES

ID	*Template*
1	*Perolehan suara terbanyak pada* **{{event}}** *diraih* **{{entity}}** *dengan* ***{{value}}*** *suara.* (The largest number of votes in **{{event}}** went to **{{entity}}** with **{{value}}** votes) **Conditions** Value_type = *Jumlah Suara* (Number of candidate's votes) Rank = 1 Entity Type = *Pasangan Calon (*Candidate) Couple = 2
2	*Perolehan suara terbanyak pada* **{{event}}** *diraih* **{{entity}}** *dengan* **{{value}}** *persen suara*. (The largest number of votes in **{{event}}** **went** to **{{entity}}** with **{{value}}** percent) **Conditions** Value_type = *Persentase Suara* (Percentage of candidate's votes) Rank = 1 Entity Type = *Pasangan Calon* (Candidate*)* Couple = 1

An aggregated template is a combination of a template couple. For example, the result of aggregating templates from Table 1 can be seen in Figure 6.

Perolehan suara terbanyak pada **{{event}}** *diraih* **{{entity}}** *dengan* **{{value 1}}** *suara atau* **{{value 2}}** *persen suara.*
(The largest number of votes in **{{event}}** went to **{{entity}}** with **{{value1}}** votes or **{{value2}}** percent)

Figure 6. Example of Aggregated Template

Templates are generated manually and automatically. Manual template generation is used so that user can tailor what kind of sentence will be generated.

Automatic template generation modifies one or more sentences into sentences with slots and conditions. Slots are generated by identifying elements in the sentences. Element that belongs to entity is turned into an entity slot. It applies to the determinations of all the slots. The automatic generation is limited to generate single template.

Generated templates are stored into database. Stored templates are selected in microplanning module automatically by rules that are defined in subsection *D*.

B. Transformation to Facts Representation

Structured data is transformed into facts representation before it is used for generating data. The facts representation consists of entity type, entity, location type, location, value type, value, event type and event. The difference between these representations to Leppänen et al. [4] facts representation is the existence of event and event type representations to explain what election that the data is explaining.

Entity type can be filled by three following values: party, candidate and elector. An entity is filled by name of the entity type that is being mentioned. Location type contains the level of the location that is being mentioned, for instance province, regency and city. Lastly, value type explains what the value is about. Example of result of transformed data can be seen in Figure 7.

No.	Entity Type	Entity	Location Type	Location	Value Type	Value	Event Type	Event
1	Pasangan Calon	Ahok-Djarot	Provinsi	DKI Jakarta	Jumlah Suara	2.350.366	Pemilihan Gubernur	Pemilihan Gubernur DKI Jakarta 2017 Putaran 2
2	Pasangan Calon	Anies-Sandi	Provinsi	DKI Jakarta	Jumlah Suara	3.240.987	Pemilihan Gubernur	Pemilihan Gubernur DKI Jakarta 2017 Putaran 2
3	Pasangan Calon	Ahok-Djarot	Kabupaten	Kepulauan Seribu	Jumlah Suara	5.391	Pemilihan Gubernur	Pemilihan Gubernur DKI Jakarta 2017 Putaran 2
4	Pasangan Calon	Anies-Sandi	Kabupaten	Kepulauan Seribu	Jumlah Suara	8.796	Pemilihan Gubernur	Pemilihan Gubernur DKI Jakarta 2017 Putaran 2

Figure 7. Result of Transformation to Facts Representation Example

C. Document Planning Module

This module consists of two tasks. In determining content, system retrieves the query-relevant subset of data that has been transformed. In addition, system also summarizes data or produces new contents that are derived from existing data using summarization rules. An example of summarization rules can be seen in Figure 8. Candidates or parties rank is also produced as new contents by ordering content according to the values and rank is given based on the order.

New value type = *Persentase Suara* (percentage of candidate's votes)
Operation = *Jumlah Suara* (number of candidate's votes) / *Jumlah Suara Sah* (number of valid votes)

Figure 8. Summarization Rule Example

After contents are determined, they are ordered based on entity type, location type and rank. If content has entity type that is the same as the news focus, it will be mentioned earlier. If content has higher level of region – for example province is higher than regency – it will be mentioned earlier. Contents with higher rank also will also be mentioned earlier in the news. After contents are ordered, they are divided into groups without changing the order of contents. A group has contents with the same entity type and location type, and similar value types. A group is expected to have similar data.

The output of this module is a list of groups consisting contents. A group describes a paragraph candidate and content describes a sentence candidate.

D. Microplanning Module

This module consists of three tasks. The first task is template selection, a limited form of lexicalization. In this task, system chooses template that is suitable for each contents. There are three different strategies in selecting template. In the first strategy, if two consecutive contents in the same group can use a template couple, they will likely be

mapped to those template couple in order to make aggregation happens. In the second strategy, if two contents in the same group have the same location and value type, they will likely be mapped to the same template. The last strategy, if the other two strategies do not correspond, template is selected by picking the least selected template in database that corresponds to content. Templates that is chosen is retrieved as sentence candidate.

In aggregation task, two sentence candidates are combined into one. The purpose of this task is to reduce chopped, short and repetitive sentences. There are two ways in aggregating sentences: using aggregation template and simple conjunctions. If two consecutive sentence candidates are a template couple, they would be modified into aggregated template that can be seen in Figure 6 as an example.

If two consecutive sentence candidates have the same template, they would be modified into a single sentence that is aggregated using simple intra-sentences conjunctions (additive and contrastive). Additive conjunctions (e.g. and) are chosen when two consecutive sentences have the same entity, value type and location, while contrastive conjunctions (e.g. whereas) are chosen when they have different entities or locations. An example of aggregation using simple intra-sentences conjunctions can be seen in Figure 9. Aggregation is also used to combine two paragraph candidates into one if a paragraph candidate only has a single sentence candidate.

Sentence 1	:	*Di Kepulauan Seribu, Anies-Sandi memeroleh suara sebanyak 8.796.* (In Kepulauan Seribu, Anies-Sandi received 8,796 votes)
Sentence 2	:	*Di Kepulauan Seribu, Ahok-Djarot memeroleh suara sebanyak 5.391.* (In Kepulauan Seribu, Ahok-Djarot received 5,391 votes)
Aggregation Result	:	*Di Kepulauan Seribu, Anies-Sandi memeroleh suara sebanyak 8.796,* ***sedangkan*** *Ahok-Djarot memeroleh suara sebanyak 5.391.* (In Kepulauan Seribu, Anies-Sandi received 8,796 votes, **whereas** Ahok-Djarot received 5,391 votes)

Figure 9. Aggregation Example using Simple Conjunctions

In REG (referring expression generation) task, if two consecutive contents have the same entity, the latter's entity turns into a referring expression. A referring expression can be in the form of the entity fact, for example the alias of a candidate. The least selected entity fact is likely chosen as referring expression. Furthermore, a referring expression can be in the form of simple pronouns or any other identifiers for the entity if there is no defined entity fact. The example of REG can be seen in Figure 10.

Sentence 1	:	*Ahok-Djarot memeroleh 2.350.366 suara di DKI Jakarta.* (Ahoh-Djarot received 2,350,366 votes in DKI Jakarta)
Sentence 2	:	*Di Kepulauan Seribu, Ahok-Djarot memeroleh suara sebanyak 5.391.* (In Kepulauan Seribu, Ahok-Djarot received 5,391 votes)
REG Result	:	*Ahok-Djarot memeroleh 2.350.366 suara di DKI Jakarta. Di Kepulauan Seribu,* ***Pasangan Calon tersebut*** *memeroleh suara sebanyak 5.391.* (Ahok-Djarot received 2,350,366 votes in DKI Jakarta. In Kepulauan Seribu, **this pair of candidates** received 5,391 votes)

Figure 10. Example of REG

The output of this module is a list of groups containing contents. Contents are mapped to templates with some of the templates are aggregated or consisting referring expression.

E. Realization Module

In this module, there are two tasks. In the linguistic realization task, slots in templates or sentence candidates are filled with relevant data. Numbers in sentences are written with proper number format in Indonesian. In addition, a sentence should start with capitalized alphabet and end with full stop. Excessive spaces are removed.

In the task structure realization, sentences in the same group are structured into a paragraph. All paragraphs then form a single news article. Two sentences are separated by a space and two paragraphs are separated by two lines. The output of this module is a news article.

The automatic news generator is a web-based application implemented using Python and Flask web framework and using MySQL as its DBMS to store input data and templates. This web application consists of two pages, the user-query page and generated news page.

The structured data used as input to generate news is mostly taken from Indonesian 2017 elections committee site (pilkada2017.kpu.go.id). Data is scraped and stored in database before it is transformed into facts representation.

The generated news can be saved in a text file. One of the generation results can be seen in Figure 11.

Bandung - Perolehan suara terbanyak pada Pemilihan Gubernur DKI Jakarta 2017 Putaran Kedua diraih pasangan calon Anies Baswedan, Ph.D. dan Sandiaga Salahuddin Uno dengan 3.240.987 atau 57,96 persen suara. Di DKI Jakarta, pasangan Ir. Basuki Tjahaja Purnama, M.M. dan Drs. H. Djarot Saiful Hidayat, M.S. mendapat 42,04 persen atau 2.350.366 suara.

Di Jakarta Barat, pasangan Anies-Sandi memperoleh 52,82 persen atau 684.980 suara, sedangkan pasangan Ahok-Djarot memperoleh 47,18 persen atau 611.759 suara. Di Jakarta Pusat, pasangan Anies-Sandi memperoleh suara 333.033 atau 57,77 persen, sedangkan pasangan Ahok-Djarot memperoleh suara 243.416 atau 42,23 persen. Di Jakarta Selatan, pasangan Anies-Sandi mendapat suara 754.665 atau 62,15 persen, sedangkan pasangan Ahok-Djarot mendapat suara 459.639 atau 37,85 persen. Di Jakarta Timur, pasangan Anies-Sandi 993.173 suara atau 61,87 persen, sedangkan pasangan Ahok-Djarot 612.093 suara atau 38,13 persen. Di Jakarta Utara, pasangan Anies-Sandi memperoleh total 52,73 persen suara atau 466.340 suara, sedangkan pasangan Ahok-Djarot memperoleh total 47,27 persen suara atau 418.068 suara.. Pasangan Anies-Sandi meraih 62,00 suara atau 62,00 suara di Kepulauan Seribu, sedangkan pasangan Ahok-Djarot meraih 38,00 suara atau 38,00 suara.

(Bandung – The largest number of votes in DKI Jakarta 2017 Governor Election Second Round was achieved by Anies Baswedan, Ph.D. and Sandiaga Salahuddin Uno with 3,240,987 or 57.96 percent of the votes. In DKI Jakarta, Ir. Basuki Tjahaja Purnama, M.M. and Drs. H. Djarot Saiful Hidayat, M.S. got 42.04 percent or 2,350,366 votes.

In West Jakarta, Anies-Sandi earned 52.82 percent or 684,980 votes, while Ahok-Djarot gained 47.18 percent or 611,759 votes. In Central Jakarta, Anies-Sandi earned 333,033 votes or 57.77 percent, while Ahok-Djarot got 243,416 votes or 42.23 percent. In South Jakarta, Anies-Sandi got 754,665 votes or 62.15 percent, while Ahok-Djarot got 459,639 votes or 37.85 percent. In East Jakarta, Anies-Sandi got 993,173 votes or 61.87 percent, while Ahok-Djarot got 612,093 votes or 38.13 percent. In North Jakarta, Anies-Sandi earned a total of 52.73 percent of the vote or 466,340 votes, while Ahok-Djarot earned a total of 47.27 percent of the vote or 418,068 votes. Anies-Sandi won 62,00 votes or 62,00 votes in Kepulauan Seribu, while Ahok-Djarot won 38,00 votes or 38,00 votes.)

Figure 11. News generation result

IV. Evaluation and Discussion

News articles generated by the system were evaluated manually by human evaluator. We asked 40 respondents to read and assess three news articles. Respondents consisted of 15 experts (journalists or linguists) and 25 active non-expert news readers. The topics of the evaluated news were DKI Jakarta 2017 governor election cycle 2, Bekasi 2017 regent election and Gorontalo 2017 governor election.

The purpose of the evaluation is to see if the generated news satisfies the correct Indonesian news characteristics. Indonesian news must have 5W + 1H elements [7]. It has to have a straightforward and not ambiguous language, should contain general, common and simple words, and should have systematic message delivery. The content should be delivered in short and dense sentences. News should sound neutral and objective. Active, short and positive sentences are much preferred. Furthermore, news should be interesting to read [8].

All of the characteristics mentioned above were used as the metric of evaluation for the human evaluator. Each respondent scored 1 – 4 for every news characteristic with 1 representing the lowest score and 4 representing the highest score. The result of the news evaluation can be seen in Table 2.

Table 2. News Evaluation Result

No.	Criteria	Average of Experts Score (Out of 4)	Average of Non-Experts Score (Out of 4)	Average of Overall Score (Out of 4)
1	News has 5W + 1H elements	2.933	2.947	2.945
2	News has straightforward and not ambiguous language	3.444	3.493	3.475
3	News uses common, simple and general words	3.178	3.427	3.333
4	News is short and dense	3.111	3.400	3.292
5	News is delivered systematically	3.244	3.307	3.283
6	News is neutral and objective	3.622	3.653	3.642
7	News uses active sentences	3.533	3.360	3.425
8	News uses short sentences	3.244	3.333	3.300
9	News uses positive language	3.622	3.507	3.550
10	News is interesting	2.756	2.627	2.675
Total Average		**3.269**	**3.314**	**3.292**

There are several issues of the generated news that caused the score subtraction by the evaluators. Respondents found that most of the news only consisted of 4 out of 6 5W+1H elements (who, when, where and what). Respondents also found that most of the news had too many unnecessary repetitions which result in the score reduction regarding short sentences and news. For example, in Figure 12, the location 'Bekasi' was repeated in the second sentence when it was clear that the previous sentence had mentioned 'Bekasi'. It should not have been mentioned in the second sentence unless both sentences talked about different locations. The repetition of the location happened because of the rule that a template should have a location slot.

*Pasangan Calon tersebut menyusul di posisi kedua dengan 309.410 suara, Obon di **Bekasi**. Di **Bekasi**, Tabroni dan Bambang Sumaryono mendapat perolehan 17,58 persen.* (This pair of candidates followed in second place with 309,410 votes in **Bekasi**. In **Bekasi**, Obon Tabroni and Bambang Sumaryono received 17.58 percent.)

Figure 12. Unnecessary Repetition in Mentioning Location

Another issue regarding repetition can be seen in Figure 13. The underlined words were not supposed to be written because it had been stated before. This repetition happened because the sentence was the result of an aggregation using a simple conjunction that concatenates two sentences directly using a conjunction without erasing the shared elements of sentences.

*Hasil akhir rekapitulasi penghitungan suara tingkat Provinsi DKI Jakarta menetapkan pasangan Anies Baswedan, Ph.D. dan Sandiaga Salahuddin Uno memeroleh 3.240.987 suara dan 57,96 persen, sedangkan **hasil akhir rekapitulasi penghitungan suara tingkat Provinsi DKI Jakarta** **menetapkan** Ahok-Djarot memeroleh 2.350.366 suara dan 42,04 persen.* (The final result of the vote recapitulation in DKI Jakarta Province determined that Anies Baswedan, Ph.D. and Sandiaga Salahuddin Uno received 3,240,987 votes and 57.96 percent, while **the final result of the vote recapitulation in DKI Jakarta Province determined** Ahok-Djarot to got 2,350,366 votes and 42.04 percent.)

Figure 13. Unnecessary Words Repetition

Some sentences were found to be too long. Those sentences were the result of aggregated sentences that had the same template. For example, the sentence in Figure 14 was considered too long. It was the combination of five different sentence candidates having the same template. In aggregating sentences using simple conjunction, if consecutive sentences have the same template, they are concatenated directly. Some shared elements of the sentence in Figure 14 can be removed as well.

Partisipasi pemilih di Jakarta Barat dengan persentase pemilih 76,40 persen warga yang menggunakan hak memilih, partisipasi pemilih di Jakarta Pusat dengan persentase pemilih 76,00 persen warga yang menggunakan hak memilih, partisipasi pemilih di Jakarta Selatan dengan persentase pemilih 75,00 persen warga yang menggunakan hak memilih, partisipasi pemilih di Jakarta Timur dengan persentase pemilih 78,66 persen warga yang menggunakan hak memilih, partisipasi pemilih di Jakarta Utara dengan persentase pemilih 77,20 persen warga yang menggunakan hak memilih, sedangkan partisipasi pemilih di Kepulauan Seribu dengan persentase pemilih 80,35 persen warga yang menggunakan hak memilih.

(Voter participation in West Jakarta with percentage of voters 76.40 percent of citizens using voting rights, voter participation in Central Jakarta with percentage of voters 76.00 percent of citizens using voting rights, voter participation in South Jakarta with percentage of voters 75.00 percent of citizens who using voting rights, voter participation in East Jakarta with voter percentage 78.66 percent of citizens using voting rights, voter participation in North Jakarta with percentage of voters 77.20 percent of citizens using voting rights, while voter participation in Kepulauan Seribu with percentage of voters 80,35 percent of citizens are using the right to vote.)

Figure 14. Unnecessary Long Sentences

Some entities (e.g. candidate names) were mentioned too long. For example, one of the pair candidates mentioned was 'Drs. H. Rusli Habibie, M.Ap and Dr. Drs. HI. Idris Rahim, MM', which included their academic degrees. The academic degrees can be removed. Another alternative is using their alias or nickname instead of their full name.

Some respondents pointed out that readers would not really care about the exact number because it is not interesting and considered too long to read. Therefore, numbers do not need to be written exactly with their decimals. They can be either rounded or summarized. For example, 57.96 percent can be written as 58 percent and 3,240,987 votes can be written as above 3 million votes or around 3.2 million votes.

The least scored news criteria was regarding the interestingness of the news. Respondents felt that there were too many numbers and the sentences were boring and not inviting. There should be more interesting facts in the news. Sentences or templates should be more varied and engaging to read.

The overall score showed a quiet high score. It was because the generated templates used for producing news were taken from published news articles that were written by professional journalists in well-known media. Sentences that were used for templates usually had fulfilled the news criteria in Table 2.

V. Conclusion

Automatic Indonesian Municipal Elections news generator is built using natural language generation architecture which consists of seven tasks, i.e. content determination, document structuring, template selection, aggregation, referring expression generation, linguistic realization, and structure realization. Template-based approach is used; thus, the template generation process is included. This system also adapts Leppänen et al. [4] facts representation.

Evaluation of the generated news was conducted using Indonesian news characteristics as the evaluation metrics. The average score of the generated news evaluated by 40 respondents (experts and non-experts) was 3.292.

To improve performance, an identifier for removable shared elements of combined sentences can be built to reduce repetitions. There can also be an academic degree remover. A number abbreviator can be applied as well. Template generation and selection can be modified to produce more interesting and varying sentences.

References

[1] Automated Insights. n.d. Wordsmith. https://automatedinsights.com/wordsmith. Accessed: 2018-06-08

[2] M. Carlson. 2015. The Robotic Reporter. *Digital Journalism,* 3(3):416–431

[3] L. Nesterenko. 2016. Building a system for stock news generation in Russian. In C Gardent and A Gangemi, editors, *Proceedings. of the 2nd International Workshop on Natural Language Generation and the SemanticWeb (WebNLG2016)*. 37 – 40.

[4] L. Leppanen, M. Munezero, M. Granroth-Wilding and H. Toivone. 2017. Data-driven news generation for automated journalism. *Proceedings of The 10th International Natural Language Generation conference*, 188–197

[5] E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.

[6] KPU. 2017. Pilkada DKI Jakarta Putaran Kedua. https://pilkada2017.kpu.go.id/hasil/2/t1/dki_jakarta. Accessed: 2018-06-08

[7] A. S. M. Romli. 2003. *Jurnalistik Praktis Untuk Pemula*. Bandung: PT Remaja Rosdakarya.

[8] Yurnaldi. n.d. *Jurnalistik Siap Pakai*. Padang: Angkasa Raya