**Automatic Abstractive Summarization Task**

**Afrida Helen**

Politeknik Elektronika Negeri Surabaya

Kampus PENS, jln Raya ITS, Sukolilo Surabaya

helen@pens.ac.id

**Abstract**

Understanding the contents of numerous documents requires strenuous effort. While manually reading the summary or abstract is one way, automatic summarization offers more efficient way in doing so. The current research in automatic summarization focuses on the statistical method and the Natural Processing Language (NLP) method.

Statistical method produce Extractive summary that the summaries consist of independent sentences considered important content of document. Unfortunately, the coherence of the summary is poor. Besides that, the Natural Processing Language expected can produces summary where sentences in summary should not be taken from sentences in the document, but come from the person making the summary. So, the summaries closed to human-summary, coherent and well structured.

This study discusses the tasks of generating summary. The conclusion is we can find that there are still opportunities to develop better outcomes that are better coherence and better accuracy.

**Key word:** Summarization,Abstactive, Extractive, Statistic, Natural Processing Language.

**1. INTRODUCTION**

People are now depending on the internet in searching for documents. By using existing tools, such as Google, Yahoo, Bing and so forth, the documents are easy to obtain, in large quantities, and coming from various sources. To understand the content of the documents quickly, the readers should read the summary (or abstract) of the documents. For summary usually contain only the important sentences in short form, and represent the contents of the document as a whole.

The summary system was built by Luhn [1] in the late 1950s. He built abstracts of scientific papers, which are placed on the top position in the scientific paper format, so that the readers have a choice to understand the contents of a paper through the Abstract. The Abstract is written by the author manually, has few amount sentences and contains only important sentence from the paper.

The problem occours when the amount of the documents is numerous. The readers obviously need more times to read and understand, despite of the summary system in each document. Therefore, the need of more advance summary system increases.

An automated summary system is one if the advance. It is generated from a collection of important sentences from a paper, whereas important sentences are constructed from important word or phrase. The system starts from marking the important words with the degree of occurrence of words or phrase in entire of papers. Sentences are ranked based on the frequency of word occurrences. Top rank of the sentence list becomes summary sentence. Then the technique to choose important sentence evolve, the characteristic of sentence involves the position of the sentence, the length of the sentence and so on. These characteristics make the accuracy of summary increasily. The common processes of summarization as shown in Fig. 1. However, this technique is low in coherence because the content of each sentence in the summary has no relation.

This weakness triggers the researchers to develop new method like human-summaries. The main characteristic of human-summary sentences are having good relationship with each other, good sentence structure, and using rhetorical move from the beginning to the end of the summary. Researchers developed the abstractive concepts that resemble human-summary. This study looks at new opportunities to make Abstractive summaries results more coherent. Several tasks of Abstactive summarization will be discussed in Section 2, followed by the opportunity in developing Abstractive Summary in Section 3. Section 4 will discuss about the opportunity of the research development. Finally, the last Section will explain the conclusion of this study.



Fig 1. The Process for Generating Summarization Automatically

**2. ABSTRACTIVE SUMMARIZATION**

Technology evolves along with needs. Similarly, the need for better summary results has led researchers to investigate the possibilities for automatic summary. Summaries made by humans have abstraction values, where sentences in summary are new sentences that essentially convey what is contained in the original document. The automatically generated abstractive summaries are expected to be better and closer to the human-built summary.

Abstractive summary requires a deeper analysis of the text and the ability to generate new sentences, which provide an obvious advantage to improve the accuracy of a summary, reducing its redundancy and keeping a good compression rate. The tasks for generating abstractive summaries [2] are 1) sentence Compression that removes peripheral information from a sentence to shorten summary, 2) sentence Fusion,  that merge information from multiple sentences and  reduces redundancy in summary and 3) reorganization Sentences to make the summary coherent. Most Abstractive summary task uses Natural Language Processing. The natural language is a task for parsing of sentences in the text. Parsing is identification the Parts of Speech of each word and the grammatical relations between the words in the sentence. The output of parsing is usually a tree of grammatical relations and dependencies between the words in the text. These trees are called dependency trees.

**2.1 Sentence Compression**

Sentence compression can be broadly described as the task of creating a grammatical summary of a single sentence with minimal information loss. The task requires a quantity called inter-phrase dependency strength. In the training process, original sentences are parsed. The number of tokens is counted for each pair of phrases, and connected with each other by a dependency path of certain length. The statistics is used to estimate the interphrase dependency strength required in the sentence compression process.

Prior tasks mostly use Decision Tree algorithm [3], random forest, and gradient boosting algorithms to analyse the data. Madanapali’s {1} research used Intersection Algorithm to align paired sentences and a swallow parser to combine the sentences. The sentence was maped into the structure of the predicate, get news content of phrase, and compare the predicate. Phrases that contain general information were selected, sorted, and added with some entities. Those were combined and arranged to generate the Summary. The results could be concluded using a tree that can improve the quality of the language of the summary significantly, and simultaneously minimize repetition. Unfortunatelly this approach did not involve the context of text when exchange between sentences.

Furthermore, Genest and G. Lapalme [2] proposed a short summary and a good abstract from several articles in the same topic. This scheme was the extraction of information based on heuristic content selection and one or more rules to generate sentences. Each abstraction is tailored to the topic or sub category. When the rule was raised, multiple verbs and nouns that have the same meaning were omitted and the position of the rule was identified. The extraction process found several candidate rules on each topic of each category. Based on information extraction module, the content module selected the best rule candidate from each category and passed it to the summary generation module. The best candidates of the content were selected using a single rule that extracted and adapted to one or two categories of rule, and then updated the rule to construct a summary sentence. This study had the potential to make a summary with more information. However, the method was purely using a rule written manually that takes a lot of time.

Another research was conducted by Trevor C and Mirella L [4]. The task was simply deleting word and rewrite sentence using additional such as, reording, insertion. They developed grammar rule for given source to set of possible output. Each rule was created by applying a series of grammar rules. where each rule matched a fragment of the source and creates a fragment of the target tree. A rule in the grammar consisted of a pair of elementary trees and a mapping between the variables in both trees. A derivation was a sequence of rules yielding a target tree with no remaining variables. Each grammar rule could assign a weigh. These weigh are learnt in discriminative training, find to set of releted sentences target for agiven source sentence and create output.

**2.2 Sentence Fussion**

We cannot separate the task of compression and fusion when constructing a summary. Both of them have to be done in order to generate a summary.

When a sentence has been compressed which generates two or more sentences that have the same intent, the fusion technique is required to combine the two sentences or more and then discard the repetitive words. The heuristic technique used is very closed to a number of rules because the most appropriate approach is to find a number of rules based on existing examples. Similarly, the fussion task, the rule is extracted using the language pattern, then is used to identify the piece of text. The Dependency trees are used in this task. They play a prominent role in sentence fusion. Sentences, which will be fused, are represented at first as dependency trees. These trees are merged in to a tree and converted to a sentence that is known as the fused sentence.

However, this study has not yielded satisfactory results. S. M. Harabagiu and F. Lacatusu, [5] built a summary of several news documents with the same topic and call GISTEXTER. They used Text snippets to generate information coherence using a summary algorithm. Significant result of the research was the high level of coherence of the summary result. On the other hand, this research could not work with a large number of documents, only in one single source.



Fig 2. The architecture of Gistexter [5]

Figure 2, show the architecture of Gistexter that is developed by Harabagieu et al. This figure is devided into three tasks. The first task is a common process for building Abstract by Human. This Abstract will be a corpus and will be used for learning. Second task is multi document summarization process. The process can be distinguished, 1) if the input document has not a template yet in corpus, and 2) if the input document has a template in corpus. The process continues to extract the sentences to build a summary. Otherwise, the System automatically generate an ad-hoc template to acquire rules. The third task combines the rule of linguistc extraction pattern with co-reference knowledge to produce a good quality of summary. Texts snipped are used to generate information coherence using a summary algorithm. Significantly development of this research is the coherence highly enough. On the other hand, this research only works in a single source document, could not handle if the source in multi document.

 Furthermore, Barzilay [6] presented research on the automatic fusion of same document topic. The method for summarizing was a specific type of input: 1) news articles presenting different descriptions of the same event, 2) a content planner selects and orders propositions from an underlying knowledge base to form text content, 3) a sentence planner determines how to combine propositions into a single sentence, and a sentence generator realizes each set of combined propositions as a sentence, mapping and building syntactic structure. The content planner found an intersection of phrases by comparing the predicate-argument structures. This process selected the phrases that able to mention the common information of the topic, order them, and augment them with information needed for clarification. The next step of generating sentence begun with phrases. The task was to produce fluent sentences that combine these phrases by arranging them in new contexts. In this process, new grammatical constraints may be imposed and paraphrasing may be needed. Redundant statement in a summary is selected by one sentence from the set of similar sentences. Therefore, need to intersect the topic sentences to identify the common phrases and then generate a new sentence. Phrases, which were produced by topic intersection, will form the content of the generated summary. Then, matching the fact was done to identify similarities between phrases instead of identifying words. If paraphrasing rules are known, the predicate-argument structure of the sentences can be compared and common parts are found. Paraphrasing pattern is obtained from studying corpus then used for intersection algorithm.

**2.3 Reorganization and Revision**

 Reorganizing and Revising a sentence needs to be done for the abstractive summary to obtain coherency. Jing and McKeown (7) found that human summarization can be traced back to cut-and-paste operations of a text and proposed a revision method consisting of sentence reduction and combination modules with a sentence extraction part.

Hideki Tanaka et al [8] method did not use the coreference relation of noun phrases (NPs), but rather insertion and substitution of the phrases to modify the same head chunk in lead and other sentences. It addressed the problem of revising the lead sentence in a news text to increase the amount of the key information. For analyzing, the method suggested to devise the lead sentence revision algorithm and present the outline. The syntactical analyzed are, 1) Trigger search for the same chunks in the lead and body sentences, 2) Phrase alignment that identify the maximum phrases of each trigger of which phrases are aligned according to a similarity metric, 3) Substitution if a body phrase has a corresponding phrase in the lead. The body phrase was richer in information, so the method substituted the body phrase for the lead phrase, and finally 4) Insertion if a body phrase has no counterpart in the lead that the phrase is floating.

The method inserts and substitutes any type of phrase that modifies the trigger and therefore had no limitation in syntactic type. Although NP elaborates, there are other useful syntactic types for revision. Khodra.M.L et al [9] evoked a summary with 15 templates for sentences from one scientific paper, and extracted seven features using statistical methods. They proposed summary coherence by substituting certain words such as subject, active verb, passive verb, phrase substitution, and discarding unimportant phrases called Surface Repair. Manually evaluated, readers are quite satisfied with the results of the summary.

**3. FUTURE RESEARCH**

Automatic summarization becomes a necessity when we are faced with pages or multiple documents or documents from multiple sources, so that the important points of the documents can be understood easily and quickly. The abstractive summary approach is expected to provide solutions or to address these issues. Researches conducted by researchers in the field of text summarization have not provided sufficient results, thus development is highly needed. From several studies discussed in the previous Section, this paper proposes some potential for developing a summary research in order to get a better summary result.

1. News documents have multiple formats and sources, it is difficult to find the standard structure for writing a document primarily for news documents. In a news document, the emphasis of important information that should be a candidate filler summary spreads throughout the document. Supposedly, a hybrid approach between extractive and abstraction methods is expected to produce a better summary. Detecting key words from within the document as a whole (if summaries are raised for only one document) is very appropriate. The tfidf method can be used to determine the words that become keywords. In addition, the selected word should not be limited only to the verb, but also to the noun. In accordance with the rules of language, nouns can be divided into two human objects and non-human objects, so when parsing the sentence, it would be better if noun is separated (like a proper noun with no prope noun). This technique can be used to find the relation (co-reference) between sentences.
2. Documents not only news that has a free format but also scientific document that is arranged in a section order. Scientific documents have a clear structure [10]. The main users of scientific documents, such as students or researchers, need particular summaries of them. Although some researchers have developed a system of summaries for scientific papers, it is still in the form of extractive summaries.
3. Similar to a news documents, in scientific documents, the key words that become candidate fillers of the summary also spread throughout in the document. However, a prior study [11] used Sections and Titles for the classification of rhetorical phrases as an approach. As a result, the key phrases relevant to summaries can be identified based on the sections presented in the document. So the next need to add abstraction techniques using natural language that makes the summary becomes more coherent and has a good structure.

**4. DISCUSSION**

The development of Abstractive Summary is a challenge for both news and scientific documents. Some tasks like compression, fussion and restructure is expected to make Abstractive Summary similar to the human built summary, althought it is still hard to recognize the human language structure. The authors send their massage in sentences in the order of phrase. But to order a phrase depend on the author’s style. This is one of obstacle to improve summary, so many structure of phrase in a sentence. So that structure-based approach is intresting to investigate. The constraint is there is no standard of the phrase structured in a sentence. The uncertain structure push researchers working hard to produce better summaries. Some methods have developed some sorts of building ontology of topics or domains, building graphs, templates, developing rules, the intention is to find significant sentences to fill a summary.

Another work is to combine sentences with the same content and then add some text to make the summary more coherent and structured. Beside that the movement the content of the sentence in summary flows in accordance with the sequence of information in the document, so the structural approach is expected to achieve a better summary.

**5. CONCLUSION**

Automatic text summaries have actually been done since the late fifties, and then become unpopular because this task difficult enough to do. But since the Internet began to be used to search the desired documents, the number of documents becomes exponentially increasing, so the need a system that can summarize the documents becomes more crucial. This phenomenon makes researchers in the field of text summarization become interested to develop a summary for better and closer to a summary made by humans, which called Abstractive summary. The challenge of this research is to improve the coherence and acuracy of the summaries. The summary content represents the contents of the document, and minimizes the repetition of sentences. This paper conducts a study of Abstractive summary that have been done by previous researchers. Each of the proposed approaches has the strengths and weaknesses. Therefore, new approaches are still needed to make the summary better like human built summaries.

**References**

1. Luhn, H.P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development.
2. PE Genest, [G Lapalme](https://scholar.google.co.id/citations?user=0fRlG8oAAAAJ&hl=id&oi=sra). (2011), Framework for Abstractive Summarization using text-to-text generation. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 64–73, Portland, Oregon, 24 June 2011. Association for Computational Linguistics (ACL)
3. Madanapalli, Andhra Pradesh, (2008), Knowledge Extraction Using Rule Based Decision Tree Approach. International Journal of Computer Scienceand Network Security IJCSNS, VOL.8 No.7, India.
4. Trevor Cohn and Mirella Lapata, (2008), Sentence Compression Beyond Word Deletion Proceedings of the 22nd International Conference on Computational Linguistics (Coling), pages 137–144 Manchester.
5. S. M. Harabagiu and F. Lacatusu, [2002], Generating single and multi-document summaries with gistexter," in Document Understanding Conferences (MUC)
6. R. Barzilay et al, [2003] ("Information fusion in  the context of multi-document summarization," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 550- 557)
7. Jing, H., & McKeown, K. R. (2000). Cut and paste based text summarization. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp. 178-185). Association for Computational Linguistics.
8. Hideki Tanaka, Akinori Kinoshita, Takeshi Kobayakawa, Tadashi Kumano, and Naoto Kato. (2009). Syntax- driven sentence revision for broadcast news summa- rization. In *Proceedings of the 2009 Workshop on Lan- guage Generation and Summarisation*, UCNLG+Sum ’09, pages 39–47, Stroudsburg, PA, USA. Association for Computational Linguistics.
9. Masayu Leylia Khodra, Dwi Hendratmo Widyantoro, E. Aminudin Aziz Bambang Riyanto Trilaksono, (2012), Automatic Tailored Multi Rhetorical Document Profile and Summary Specification, Journal of ICT Research and Applications. Published by ITB Journal Publisher, Vol. 6, No. 3, 2012, pp 220-239, ISSN: 1978-3086, DOI: 10.5614.
10. Afrida Helen, Ayu Purwarianti, Dwi Hendratmo Widyantoro (2014), Extraction and Classification of rhetorical sentences of experimental technical paper based on section class, 2nd International Conference on Information and Communication Technology (ICoICT), 2014, Date of Conference: 28-30 May 2014, IEEE *Xplore*: DOI: [10.1109/ICoICT.2014.6914099](https://doi.org/10.1109/ICoICT.2014.6914099%22%20%5Ct%20%22_blank)
11. Afrida Helen, Ayu Purwarianti, Dwi Hendratmo Widyantoro, (2015), Rhetorical Sentences Classification Based on Section Class and Title of Paper for Experimental Technical Papers, Journal of ICT Research and Applications. Published by ITB Journal Publisher, Vol. 9, No. 3, 2015, pp. 288-310, ISSN: 2337-5787, DOI: 10.5614/