Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 11–19

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

# Automatic Debate Text Summarization in Online Debate Forum

Alan Darmasaputra Chowanda[a], Albert Richard Sanyoto[a], Derwin Suhartono[a] *, Criscentia Jessica Setiadi[b]

[a] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K.H. Syahdan No.9 Kemanggisan, Jakarta 11480, Indonesia*

[b] *English Department, Faculty of Humanities, Bina Nusantara University, Jl. Kemanggisan Ilir III No.45 Palmerah, Jakarta 11480, Indonesia*

**Abstract**

The goal of this research is to create a system that can generate summaries from online debate forum by using abstractive technique. This research is based on the point-based summarization technique, where a point is a verb and its syntactic arguments. The point is extracted based on the dependency parse and the syntactic frame. Our proposed system implements the system in three different modules: point extraction, point curation, and summary generation. The system also includes the stance of statement to improve performance of the summarizer. We use ROUGE metrics to evaluate this system. The results of this research are our proposed system performs the best in terms of precision and gets the best f-score after the summaries are preprocessed. The proposed system increases the ROUGE-1 score by 8.99% compared to the point-based summarization and produces 15.84% increase compared to the baseline summarization system.

*Keywords:* automatic summarization; debate; point-based summarization

## 1. Introduction

The amount of data produced in the Internet is increasing. There are 2.4 million Google search queries, 347,222 new tweets on Twitter, and 701,389 Facebook logins for every 60 seconds [1]. It is easy to be overwhelmed by such

* Corresponding author. Tel.: +6221 5345830 ext 2188; fax: +6221 5302244.
*E-mail address:* dsuhartono@binus.edu

10.1016/j.procs.2017.10.003

enormous traffic and data that are available on the Internet. Therefore, it will be helpful if there is a tool that exists to help digesting such large amount of information. An example of such tool is a system that can automatically generate a summary from a document, which contains the brief and short version of the original document.

One of the platform where users could present their ideas to other is in online debate forums. This online media provides a large collection of opinion on various topics. In an ideological two-sided debate, one side will stand for the issues, and the other will argue against the issues. The first usually referred to as pros, and the latter as cons. Users support their stance by cleverly stating arguments supporting their stance or opposing the other stance [2]. Because of the dynamic nature of debates and large amount of posts they contain, it is essential to generate effective summaries for them so that the readers do not need to go through the entire debate to understand them [3].

This research is based on a technique developed before [4]. The technique is a point-based summarization technique which is separated into three different modules: point extraction, point curation, and summary generation. The improvement proposed in this research is to include the stance of statement to improve performance of the summarizer. The texts that will be summarized are in English and from a single online debate forum. The data will be in form of two-sided format debate, with pro and cons stance to the topics.

## 2. Related Works

There are already several techniques developed in an attempt to summarize the arguments in a debate text. The vast majority of the techniques are extractive, which concatenates several important sentences in the original material. There are also some research concerning abstractive summarizer, which expresses the main information of the document in the words of the summary author [5].

The earliest method was proposed by Luhn [6]. This approach is based on the idea that some words in a document are descriptive to its content. Therefore, sentences containing many of such words are the important ones. Later on, improvement on this approach by using a linear combination of features to weight the sentences was developed [7]. The features used are number of times a word appears, the number of words in a sentence that also appears in the title, position of the sentence, and the number of sentences matching a list of cue words.

A topic directed sentiment analysis based summary is then developed [3]. The proposed system scores each unit on the document based on several features. There are four categories of features used, which are topic relevance, document relevance, sentiment relevance, and context relevance. Another algorithm was developed by modifying the integer linear programming system [8]. The idea is to maximize the overall score while minimizing redundancy among selected sentences. Another summarization technique is proposed [4]. This technique is different from the majority of the algorithm as it is an abstractive summarizer. The concept is based on the notion of point, which contains a verb and its syntactic arguments. The document is converted into clusters of points with the same meaning. The cluster that contains the highest number of point means it represents the document. An extract is then produced from the cluster.

## 3. Proposed Method

### 3.1. System Architecture

The system is implemented as a series of process that form a processing pipeline. Each of the process is separated into modules. The result of each module will then be transferred to the other module in the next step of the processing pipeline. The system is designed in form of a pipeline due to the computing resources and processing time required by each module.