REPORT FOR MINI-HACKATHON II
ROUND I

Le Van-Duyet[1], Vo Minh Quan[2] and Nguyen Duc Thinh[3]

Abstract. Text summarization is an application of natural language processing and is becoming more popular for its useful applications. Text summarization is a process of shortening original document and producing a summary by retaining important information of the original document. The report is to introduce this insteresting yet challenging problem.

Keywords: Text summarization, NLP

Introduction

In recent years, there has been a explosion in the amount of text data from variaties of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. There has been many research on this problem, the main purpose of this report is to have an overview to this problem and explanation of a potential method.

1. Literature Review

There are two general approaches to automatic text summarization: **extraction** and **abstraction**.

Extractive methods work by identifying an informative subset of words, phrases, or sentences in the original text to form the summary. Thus, this approach focus on how to break a document into appropriate tokens (words, phrases or sentences) and ranking these tokens.

In contrast, abstractive methods aim at using natural language generation techniques to create a summary that conveys the most critical information from the original text while being closer to what a human might express.

March, 2018.

[1] duyet.le@jvn.edu.vn

[2] quanvoglandore@jvn.edu.vn

[3] ducthinh.nguyen2015@ict.jvn.edu.vn

Even though summaries created by humans are usually not extractive, most of the summarization research today has focused on extractive summarization. Extractive methods often give better result compared to abstractive methods, it is because abstractive summarization methods meet problems such as semantic representation, inference and natural language generation which are more challenging than data extraction problems.

There are some main approaches in this domain should be considered:

(1) **Statistical Approaches**

In the early time, an important research was [9], Luhn proposed that the most frequent words represent the most important concept of the text. The main idea was to give the score to each sentence based on number of occurrences of the words and then choose the sentence with the highest score. This score is called "significance factor" reflecting the number of occurrences of significant words within a sentence and the linear distance between them. Also, ignoring non-informative very high frequency words (often known as stop words) is a necessary step.

Using **TF-IDF** for representing sentences is a common approach among the studies [8] [6] [1] The term frequency-inverse document frequency is a numerical statistic which reflects how important a word is to a document. It is often used as a weighting factor in information retrieval and text mining.

Location of text is also used to weight sentences [9] [7] [5]. The idea is that, leading several sentences or last few sentences or conclusion are considered to be more important. The sentences appear in the title are also considered to be more important.

(2) **Linguistic Approaches**

Linguistic approaches are based on considering the semantically connection between the words and trying to find the main concept by analyzing these words. **Lexical chain** [3] [14] and **WordNet** [4] are among the researches following this approach.

(3) **Machine learning Approaches**

Both **supervised** and **unsupervised** algorithms has been used for this problem. Seeing this as a classification problem, many algorithms such as Naive Bayes, decision trees, support vector machines, etc are used to classify the sentences as summary sentences and non-summary sentences [11] [15]. Clustering algorithms are also applied [13] [2]. As the fast growing of deep learning, this approach still remains followed. [12] is an example of using **deep reinforced model** for abstractive summarization.

(4) **Graph Approaches**

Graph methods which are influenced by PageRank algorithm, represent the documents as a connected graph. Sentences form the vertices of the graph and edges between the sentences indicate how similar the two sentences are.

TextRank [10] is a graph based algorithm which is applies in summarization. The similarity relation is measured as the number of common tokens between lexical representations of two sentences.

LexTank [6] is a graph approach similar to TextRank. LexRank uses cosine similarity of TF-IDF vectors as the similarity measure between two sentences.

And in this report we introduce a method based on the well known TextRank algorithm with a few enhancements.

2. Approaches

We approach this problem by implement TextRank - an unsupervised algorithm based on weighted-graphs from a paper by Mihalcea et al [10], with a few enhancements like using lemmatization instead of stemming, incorporating Part-Of-Speech tagging and Named Entity Resolution.

The basic idea implemented by a graph-based ranking model is that of "voting" or "recommendation". When one vertex links to another one, it is basically casting a vote for that other vertex.

3. Deliverables

The main deliverables for this report are:

- The source code
- Live demo at abc.com

References

[1] Rasim M. Alguliev et al. "MCMR: Maximum Coverage and Minimum Redundant Text Summarization Model". In: *Expert Syst. Appl.* 38.12 (Nov. 2011), pp. 14514–14522. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2011.05.033.

[2] Utsav Gupta Ayush Agrawal. *Extraction based approach for text summarization using k-means clustering*. Nov. 2014.

[3] Regina Barzilay and Michael Elhadad. "Using lexical chains for text summarization". In: *Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization*. 1997, pp. 10–17.

[4] William Doran et al. "Comparing Lexical Chain-based Summarisation Approaches using an Extrinsic Evaluation". In: (Apr. 2004).

[5] H. P. Edmundson and R. E. Wyllys. "Automatic Abstracting and Indexing&Mdash;Survey and Recommendations". In: *Commun. ACM* 4.5 (May 1961), pp. 226–234. ISSN: 0001-0782. DOI: 10.1145/366532.366545.

[6] Gfffdfffdnes Erkan and Dragomir R. Radev. "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization." In: *J. Artif. Intell. Res. (JAIR)* 22 (Feb. 8, 2005), pp. 457–479.

[7] Eduard Hovy and Chin-Yew Lin. "Automated Text Summarization and the SUMMARIST System". In: *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*. TIPSTER '98. Baltimore, Maryland: Association for Computational Linguistics, 1998, pp. 197–214. DOI: 10.3115/1119089.1119121.

[8] Youngjoong Ko and Jungyun Seo. "An Effective Sentence-extraction Technique Using Contextual Information and Statistical Approaches for Text Summarization". In: *Pattern Recogn. Lett.* 29.9 (July 2008), pp. 1366–1371. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2008.02.008.

[9] H. P. Luhn. "The Automatic Creation of Literature Abstracts". In: *j-IBM-JRD* 2.2 (1958), pp. 159–165. ISSN: 0018-8646 (print), 2151-8556 (electronic).

[10] Rada Mihalcea and Paul Tarau. "Textrank: Bringing order into text". In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.

[11] You Ouyang et al. "Applying Regression Models to Query-focused Multi-document Summarization". In: *Inf. Process. Manage.* 47.2 (Mar. 2011), pp. 227–237. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2010.03.005.

[12] Romain Paulus, Caiming Xiong, and Richard Socher. "A Deep Reinforced Model for Abstractive Summarization". In: *CoRR* abs/1705.04304 (2017). arXiv: 1705.04304.

[13] Zhang Pei-ying and Li Cun-he. *Automatic text summarization based on sentences clustering and extraction*. Sept. 2009.

[14] H. Gregory Silber and Kathleen F. McCoy. "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization". In: *Computational Linguistics* 28.4 (2002), pp. 487–496. DOI: 10.1162/089120102762671954.

[15] Kam-Fai Wong, Mingli Wu, and Wenjie Li. "Extractive Summarization Using Supervised and Semi-supervised Learning". In: *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*. COLING '08. Manchester, United Kingdom: Association for Computational Linguistics, 2008, pp. 985–992. ISBN: 978-1-905593-44-6.