Generating Summaries Through Unigram and Bigram: Text Summarization

Nesreen Mohammad Alsharman, WISE, Amman, Jordan
Inna V. Pivkina, NMSU, USA

ABSTRACT

This article describes a new method for generating extractive summaries directly via unigram and bigram extraction techniques. The methodology uses the selective part of speech tagging to extract significant unigrams and bigrams from a set of sentences. Extracted unigrams and bigrams along with other features are used to build a final summary. A new selective rule-based part of speech tagging system is developed that concentrates on the most important parts of speech for summarizations: noun, verb, and adjective. Other parts of speech such as prepositions, articles, adverbs, etc., play a lesser role in determining the meaning of sentences; therefore, they are not considered when choosing significant unigrams and bigrams. The proposed method is tested on two problem domains: citations and opinosis data sets. Results show that the proposed method performs better than Text-Rank, LexRank, and Edmundson summarization methods. The proposed method is general enough to summarize texts from any domain.

KEYWORDS

Abstractive Summarization, Extractive Summarization, Lex-Rank, Natural Language Processing, Part of Speech Tagging, Text-Rank, Unigram-Bigram Extraction

INTRODUCTION

The quantity of online information is growing exponentially. Summarizing information returns main ideas of documents without reading all the content of text which helps to minimize a reading time and can aid search engines to recognize relevant content. A summary can be classified as an indicative way or informative way. Indicative way can be defined as a pointer to some parts of the original document. Informative way covers all relevant information of the text (Fan et al., 2006). In both cases the most important advantage of using a summary is the reduced reading time.

There are a lot of automatic text summarization approaches in the Natural Language Processing (NLP) subfield that summarize original text. Automatic summarization is a set of techniques which generate important component summaries of original text; the summaries are expected to be at most half the size of the original text. A good automatic text summarization should provide a good summary that should cover the most important information of the original text or a cluster of text,
while being coherent, non-redundant and grammatically readable. In short, summarization must satisfy the following three important optimization properties:

1. Summaries should contain important text units that are relevant to the user;
2. Summaries should not contain multiple textual units that convey the similar information; and
3. Summaries are bounded in length (Babar and Patil, 2015).

Generally, text summarization methods can be classified into extractive and abstractive summarizations, single-document and multi-document, and supervised/unsupervised summarizations. Extractive summarization produces summaries by concatenating several sentences taken exactly as they appear in the original documents being summarized (Binwahlan et al., 2009; Garg et al., 2009). Extractive summarization selects a subset of existing words, phrases, or sentences in the original text to generate a summary. By contrast, abstractive summarization (Knight and Marcu, 2002; Riezler et al., 2003; Turner and Charniak, 2005) uses different words to describe the contents of the original documents rather than directly copying original sentences (Yao et al., 2017). Abstractive summarization uses text-to-text generation or sentences compression to build a cohesive and coherency summary without redundant information.

Depending on the number of input documents, there could be single-document summarizations and multi-document summarizations. While a single document is used as an input to create a single summary in single-document text summarization systems; multiple documents related to a single subject are used in multi-document text summarization systems.

Several different methods have been proposed in the literature for supervised approaches (Das and Martins, 2007; Pei et al., 2012), and unsupervised approaches (Erkan and Radev, 2004; Mihalcea and Tarau, 2004). The main idea of supervised approaches is given a set of training documents and their extractive summaries, the summarization process is modeled as a machine learning classification problem: sentences are classified as non-summary and summary sentences depends on the features that they possess.

The challenge of this research is to develop a single unsupervised extractive summarization method that produces a summary of any text without training data set. A unigram–bigram extractive technique with other features to build a summary is used. The unigram–bigram extraction method employs on a new selective rule-based part of speech tagging which only uses the three main categories of words: noun, verb, and adjective. The proposed method fully omits other word categories like prepositions, articles, adverbs, etc.

RELATED WORK

Interest in automatic text summarization appeared as early as in the fifties. Throughout time several dominant extractive methods for producing summarizations have emerged. The methods generate summaries that retain the most significant information of a given text. Summarization methods can be categorized into frequency-based, graph-based and machine learning-based.

Edmundson’s work (Edmundson, 1969) describes one of the earliest extractive frequency-based methods. It uses a word frequency method that calculates term frequency and inverse document frequency (TF-IDF) for each word in a text. Each sentence is assigned a weight that is based on the following features: TF-IDF frequencies of its words, the presence of cue words (words like “significant” or “conclusion”), and the skeleton of the document (whether the sentence is a title or a heading). Summary is built from the sentences with the highest weights.

Recently, unsupervised graph-based summarization methods have attracted an increasing attention of researchers (Shen and Li, 2010; Wang et al., 2008) and have been successful when compared to the other state-of-the art summarization approaches (Mihalcea, 2004). Graph-based methods use graphs
Related Content

Application Case Studies
(2018). Novel Design and the Applications of Smart-M3 Platform in the Internet of Things: Emerging Research and Opportunities (pp. 105-126).
www.irma-international.org/chapter/application-case-studies/183368

Polarity Classification of Arabic Sentiments
Mohammed N. Al-Kabi, Heider A. Wahsheh and Izzat M. Alsmadi (2016).
International Journal of Information Technology and Web Engineering (pp. 32-49).
www.irma-international.org/article/polarity-classification-of-arabic-sentiments/164470

A Dynamic and Adaptable Service Composition Architecture in the Cloud Based on a Multi-Agent System
www.irma-international.org/article/a-dynamic-and-adaptable-service-composition-architecture-in-the-cloud-based-on-a-multi-agent-system/193009

Adaptability and Adaptivity in The Generation of Web Applications
International Journal of Information Technology and Web Engineering (pp. 20-44).
www.irma-international.org/article/adaptability-adaptivity-generation-web-applications/4033

A Dynamic and Adaptable Service Composition Architecture in the Cloud Based on a Multi-Agent System
www.irma-international.org/article/a-dynamic-and-adaptable-service-composition-architecture-in-the-cloud-based-on-a-multi-agent-system/193009