Information Processing and Management 54 (2018) 145–158

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Effective aggregation of various summarization techniques

Parth Mehta*[a], Prasenjit Majumder[b]

[a] Dhirubhai Ambani Institute of Information and Communication Technology, Postal Address: IR&LP-Lab DA-IICT, Near Indroda Circle, Gandhinagar 382007, Gujarat, India

[b] Dhirubhai Ambani Institute of Information and Communication Technology, Postal Address: 4209, FB-4, DA-IICT, Near Indroda Circle, Gandhinagar 382007, Gujarat, India

ARTICLE INFO

Keywords:
Summarization
Ensemble

ABSTRACT

A large number of extractive summarization techniques have been developed in the past decade, but very few enquiries have been made as to how these differ from each other or what are the factors that actually affect these systems. Such meaningful comparison if available can be used to create a robust ensemble of these approaches, which has the possibility to consistently outperform each individual summarization system. In this work we examine the roles of three principle components of an extractive summarization technique: sentence ranking algorithm, sentence similarity metric and text representation scheme. We show that using a combination of several different sentence similarity measures, rather than only one, significantly improves performance of the resultant meta-system. Even simple ensemble techniques, when used in an informed manner, prove to be very effective in improving the overall performance and consistency of summarization systems. A statistically significant improvement of about 5% to 10% in ROUGE-1 recall was achieved by aggregating various sentence similarity measures. As opposed to this aggregation of several ranking algorithms did not show a significant improvement in ROUGE score, but even in this case the resultant meta-systems were more robust than candidate systems. The results suggest that new extractive summarization techniques should particularly focus on defining a better sentence similarity metric and use multiple sentence similarity scores and ranking algorithms in favour of a particular combination.

1. Introduction

Automatic Summarization, or reducing a text document while retaining its most important points, is not a new research area. The first notable attempt was made by Luhn (1958), which uses word frequencies to identify significant words in a given sentence. The importance of a sentence is then determined from the number of significant words it has and proximity of these words to each other. Since then the techniques for both sentence selection (extractive summarization) as well as abstract generation (abstractive summarization) have advanced a lot. Unfortunately most works reported before beginning of the new millennium, were either not reproducible due to lack of standard evaluation corpora (Brandow, Mitze, & Rau, 1995), (Kupiec, Pedersen, & Chen, 1995) or worse, they were not evaluated at all. It was only in last decade and a half that streamlined efforts were made possible due to conferences like DUC (Dang, 2005) and TAC (Owczarzak & Dang, 2011) which generated standard evaluation benchmarks for text summarization.

Numerous summarization techniques have been suggested in past that range from simple frequency based approaches such as FreqSum (Nenkova, Vanderwende, & McKeown, 2006), which rank sentences based on the frequency of its words in the document, to

* Corresponding author.
E-mail addresses: parth_me@daiict.ac.in (P. Mehta), p_majumder@daiict.ac.in (P. Majumder).

https://doi.org/10.1016/j.ipm.2017.11.002
Received 23 November 2016; Received in revised form 30 September 2017; Accepted 9 November 2017

the sophisticated sentence compression techniques, like Cohn and Lapata (2009), which also focus on rewriting sentences besides compression. The survey by Das and Martins (2007) broadly categorizes these techniques into Extractive, Abstractive and Information Fusion techniques. Extractive techniques, as the name suggests, solely rely on extracting important sentences from a document or set of documents.

As compared to that, abstract generation approaches seem more natural and closely relate to the way humans summarize the documents. But the success in abstract generation has largely been limited due to the inability of present computational techniques to generated very fluent and grammatically correct abstractive summaries, as well as their dependency on linguistic resources. Lately, as more training data becomes available and with the recent advances in the field of deep learning, some fresh attempts like Rush, Chopra, and Weston (2015) and Chopra, Auli, and Rush (2016) are now being made towards data driven abstractive summarization that have minimal dependency on linguistic inputs.

Sentence compression techniques have evolved as a bridge between sentence extraction and abstractive techniques. These are still closer to the extractive techniques, but also perform operations like phrase deletion (Knight & Marcu, 2002) to shorten the extracted sentences and sometimes use sentence fusion to combine two or more related sentences (Barzilay & McKeown, 2005). In this work we will solely focus on extractive techniques.

The survey by Nenkova and McKeown (2012) categorizes the extractive summarization system based on whether they are topic representation based or indicator representation based. A topic representation based system can vary from as simple representation as tf-idf (Radev, Jing, Styś, & Tam, 2004) or word frequency (Nenkova et al., 2006), to topic signatures (Lin & Hovy, 2000) or latent semantic indexing. One very popular topic representation based system is the Centroid based method. It depends on a representative centroid, which is computed based on the tf-idf representations of each sentence in the document. Any sentence close to the centroid is deemed to be more important. Unlike this technique which use tf-idf vectors to represent sentences, some other methods use alternate text representations like term distributions (Haghighi & Vanderwende, 2009) or Latent semantic indexing (Steinberger, 2004). Some recent attempts, like Kobayashi, Noguchi, and Yatsuka (2015) and Kågebäck, Mogren, Tahmasebi, and Dubhashi (2014), have used word embeddings based representation. As opposed to these, indicator representation approaches do not rely on extracting or interpreting topics. They instead represent a document in a way that direct ranking of sentences becomes possible. Some well known approaches of this kind are the graph based approaches. Extractive techniques like LexRank (Erkan & Radev, 2004) and TextRank (Radev et al., 2004) simply try to select the best representative set of sentences from a given document/document cluster without any modification. In such techniques each document is represented as a graph and each sentence in the document constitutes a node. The sentences are ranked using an approach similar to the popular Pagerank algorithm (Page, Brin, Motwani, & Winograd, 1999).

Most of these extractive techniques comprise of three fundamental components: (1) Sentence similarity metric, (2) Sentence ranking algorithm and (3) Text representation scheme. These components are relatively independent of each other, and their choice can drastically alter performance of a summarization system. Any new extractive technique proposes a particular combination of these components. But few inquiries have been made as to how changing these components affects the overall effectiveness of the system. This is the first question that we address in this work.

Due to a common platform provided by the likes of DUC and TAC, it is now possible to compare performance of a plethora of existing systems and create a meta-system that, at least theoretically, can outperform each individual system. In their work, Hong et al. (2014) provide a comparison of summaries generated by several standard and state-of-art techniques. They evaluate all these systems using the same ROUGE setup with a fixed set of parameters, thus making the scores comparable. ROUGE (Lin, 2004) is a family of metrics based on n-gram overlap, and has a number of parameters which determine the exact evaluation measure. The same system can obtain very different ROUGE scores when evaluated with a different set of parameters. Hong et al. (2014) highlight the fact that although many different systems have similar ROUGE scores under this common setup, the content across these are substantially different. This indicates that there is a scope for combining summaries generated by different systems and doing so can improve the coverage of resultant summary. Despite this, few attempts have been made to leverage this difference in performance of summarization systems to create an ensemble or inquire if such an ensemble will be useful at all.

2. Research objective

Following are the main questions that we attempt to address in this work:

- What effect does the individual components of an extractive summarization system have on the overall performance?
- How sensitive is a new summarization system to change in these components?
- Can the answers to above questions be leveraged to generate better ensemble system?

In this work we start by highlighting the effects of preprocessing and post processing steps on performance of the overall system. We then proceed to demonstrate the effect of variation in three principle components of an extractive summarization system: sentence similarity metric, ranking algorithm and text representation technique. Any new extractive summarization technique is proposed with a particular choice of these components as well as other pre/post processing steps. But they usually fail to provide any insights as to why this particular choice. As a matter of fact most combinations of the available sentence similarity metric and ranking algorithms would be valid and can form a new summarization system of their own. We use such variations in to create a large number of relatively similar systems and show that the originally proposed combinations don't always perform better.

Since it is not always possible to predict beforehand which combination would perform the best, we propose using multiple