Expert Systems With Applications 53 (2016) 43–56

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Concept generalization and fusion for abstractive sentence generation

Riadh Belkebir*, Ahmed Guessoum

Natural Language Processing and Machine Learning Research Group, Laboratory for Research in Artificial Intelligence, Computer Science Department, University of Science and Technology Houari Boumediene (USTHB), BP 32 El-Alia, Bab Ezzouar, Algiers 16111, Algeria

ARTICLE INFO

Keywords:
Artificial intelligence
Natural language processing
Text summarization
Concept fusion
Concept generalization
WordNet

ABSTRACT

Text summarization is either extractive or abstractive. Extractive summarization is to select the most salient pieces of information (words, phrases, and/or sentences) from a source document without adding any external information. Abstractive summarization allows an internal representation of the source document so as to produce a faithful summary of the source. In this case, external text can be inserted into the generated summary. Because of the complexity of the abstractive approach, the vast majority of work in text summarization has adopted an extractive approach.

In this work, we focus on concepts fusion and generalization, i.e. where different concepts appearing in a sentence can be replaced by one concept which covers the meanings of all of them. This is one operation that can be used as part of an abstractive text summarization system. The main goal of this contribution is to enrich the research efforts on abstractive text summarization with a novel approach that allows the generalization of sentences using semantic resources. This work should be useful in intelligent systems more generally since it introduces a means to shorten sentences by producing more general (hence abstractions of the) sentences. It could be used, for instance, to display shorter texts in applications for mobile devices. It should also improve the quality of the generated text summaries by mentioning key (general) concepts. One can think of using the approach in reasoning systems where different concepts appearing in the same context are related to one another with the aim of finding a more general representation of the concepts. This could be in the context of Goal Formulation, expert systems, scenario recognition, and cognitive reasoning more generally.

We present our methodology for the generalization and fusion of concepts that appear in sentences. This is achieved through (1) the detection and extraction of what we define as generalizable sentences and (2) the generation and reduction of the space of generalization versions. We introduce two approaches we have designed to select the best sentences from the space of generalization versions. Using four NLTK[1] corpora, the first approach estimates the “acceptability” of a given generalization version. The second approach is Machine Learning-based and uses contextual and specific features. The recall, precision and F1-score measures resulting from the evaluation of the concept generalization and fusion approach are presented.

1. Introduction

Text summarization is one of the most difficult, though promising, applications of Artificial Intelligence (AI) in general, and Natural Language Processing (NLP) more specifically. Various prestigious conferences and organizations have paid special attention to this field. One can mention the Association for the Advancement of Artificial Intelligence (AAAI[2]), the Document Understanding Conferences (DUC[3]) and, the Text Analysis Conference (TAC[4]). Various definitions of text summarization are given in the literature. Hovy and Marcu (2005) define a summary as a text which is produced from one or more texts, which contains a significant portion of the original text(s) information, and which is no longer than half of the original text(s). Mani and Maybury (1999), define the text summarization task as the process of finding the important contents in

* Corresponding author. Tel.: +213774186689.
E-mail addresses: belkebir.riadh@gmail.com (R. Belkebir), aguessoum@usthb.dz (A. Guessoum).

[1] The Natural Language Tool Kit (http://nltk.org/)

[2] http://www.aaai.org/

[3] http://duc.nist.gov/

[4] http://www.nist.gov/tac/

http://dx.doi.org/10.1016/j.eswa.2016.01.007
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

the original text and presenting them as a concise text in a predefined template.

Text summarization approaches are classified into two categories: extractive and abstractive. Extractive summarization consists in selecting the most relevant fragments (chunks of sentences, entire sentences, paragraphs) from an original document and concatenating them so as to generate a shorter text. Text summarization by abstraction is to create a new shorter document from an original one but not necessarily restricted to fragments present in the original document. In fact, new (external) pieces of information can be added to generate a summary. Currently, abstractive summarization seems to be the trend and a challenge to the community (Lloret & Palomar, 2012).

In this work, we address the problem of abstractive text summarization with a focus on the task of concept fusion and generalization. The latter can be seen as one operation among several ones that can contribute to text summarization. It is considered difficult as it requires a cognitive effort to achieve it. We are particularly interested in generalizing sentences, i.e. such that the system be able to generate from a sentence like *“Sue ate bananas, apples and potatoes”* an output like *“Sue ate fruits and vegetables”* or *“Sue ate some food”*. This task requires the use of world knowledge. In our case, we use WordNet[5] (Miller, 1995) as a source of external knowledge to generalize concepts, hence to abstract sentences.

We automatically generate the generalization and fusion of the concepts of a given sentence through a sequence of steps. The first step is to decide whether a given sentence is generalizable or not. If it is, we generate the set of possible generalizations (versions) of the sentence. The next step is to reduce the space of generalization versions. And, in order to further reduce this space and get a set of generalization versions that are acceptable in natural language, a heuristic-based and a Machine Learning-based model are proposed. Once the best generalization version is found, we generate the compressed sentence. The methodology proposed can generalize even complex sentences thanks to the dependency parsing module which is used and is described below.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 introduces the problem statement and definitions. Section 4 explains the system design. First, we tackle the problem of extraction of generalizable sentences. We then show how the space of generalization versions can be generated and then reduced. Next, we describe the heuristics we use to select acceptable versions from the space of generalization versions. The evaluation methodology and experimentation work are presented in Section 5. A running example is used in Section 6 to illustrate the whole approach. Section 7 discusses the results we have obtained and Section 8 gives a conclusion as well as a listing of some possible directions for the development of text summarization based on this work.

2. Related work

Text summarization is not a new discipline. It has actually started attracting researchers since the earliest work of Luhn (1958) in the late 1950s and Edmundson (1969) in the late 1960s. At that time, research interest was in the generation of abstracts of technical documents. This interest quickly declined due to its difficulty but revived afterwards thanks to the renewed interest the Artificial Intelligence community developed for it (Lloret & Palomar, 2012). Text summarization has been treated from different angles. In the sequel, we show that most studies have used extractive systems to tackle the problem. We present the main efforts that have been done in the literature to understand the text summarization task and present the different operations that have been used. These operations include text simplification, sentence compression and sentence fusion. We also introduce recent advances in text summarization and how abstractive text summarization is tackled nowadays. We conclude this section by giving the main contributions of our approach.

Because of the difficulty of abstractive text summarization, most studies have followed the extractive paradigm, selecting the important pieces of information from the source document verbatim (chunks of sentences, entire sentences, paragraphs), i.e., without adding any external text to the generated summary. Recently, Ferreira et al. (2013) have assessed the performance of 15 techniques for sentence scoring (in extracted texts) which are the most common in the field. The extractive approach has a lot of shortcomings in terms of the quality of the generated summary. One of these is the lack of coherence, especially due to the existence of “dangling anaphors” (Lloret & Palomar, 2012). Abstractive text summarization can theoretically solve this problem. In fact, it allows an internal representation of the source document so as to produce a faithful summary of the source, preserving thereby its readability and coherence. This approach allows the production of a summary by not only deleting (words, phrases, and/or sentences) from the source document, but by also allowing the addition to the summary of new material that was not necessarily present in the original document. An important feature of abstractive text summarization compared to the extractive one is that it generates a summary which is most of the time shorter and more informative (Jing & McKeown, 2000).

A number of studies have been done to try to understand the task of summarization. Jing (2002) has used Hidden Markov Models to decompose summaries produced by human experts. She has tried to infer whether a summary is constructed by reusing phrases from the original text, identifying these phrases and finding the positions in the original text these phrases come from. Jing and McKeown (2000) have analyzed a set of human written abstracts. They have proposed an approach that identifies the places in the source text the phrases from the abstract originate from; they also produce an aligned corpus of source texts and their corresponding summaries so these can be used to train the summarizer. In a similar study, Hasler (2007) claims that, to generate English summaries, humans copy and paste snippets from the source document after some slight modifications. These operations were classified as either atomic or complex operations. The atomic operations are deletion and insertion of words while the complex operations include replacement and reordering of words and merging of sentences. According to the evaluation they performed, 78% of abstracts were more coherent than extracts.

Various approaches have been used to tackle text summarization. Some of these are based on natural language generation. For instance, Radev and McKeown (1998), have developed a system, SUMMONS, which produces multi-document summaries of the same event by using the output of systems developed for the DARPA Message Understanding Conferences. In a similar study, Kumar, Das, Agarwal, and Rudnicky (2009) have designed a learning-based system that generates a draft report as a mix of event data and the input text document. This learning system was trained on a corpus of reports prepared by experts in the target (conference replanning) domain.

Various studies have tried to do text simplification. The focus here is on rewriting operations applied to source sentences so as to decrease the syntactic or lexical level of complexity and at the same time to preserve their meaning (Siddharthan, 2002). In this sense, Coster and Kauchak (2011) have used different simplification operations including rewording, reordering, insertion and deletion by introducing a data set that pairs Simple English Wikipedia with English Wikipedia. Similarly, Woodsend and Lapata (2011) have

[5] http://wordnet.princeton.edu/