Chapter 1
Combining Machine Learning and Natural Language Processing for Language-Specific, Multi-Lingual, and Cross-Lingual Text Summarization:
A Wide-Ranging Overview

Luca Cagliero
Politecnico di Torino, Italy

Paolo Garza
Politecnico di Torino, Italy

Moreno La Quatra
Politecnico di Torino, Italy

ABSTRACT

The recent advances in multimedia and web-based applications have eased the accessibility to large collections of textual documents. To automate the process of document analysis, the research community has put relevant efforts into extracting short summaries of the document content. However, most of the early proposed summarization methods were tailored to English-written textual corpora or to collections of documents all written in the same language. More recently, the joint efforts of the machine learning and the natural language processing communities have produced more portable and flexible solutions, which can be applied to documents written in different languages. This chapter first overviews the most relevant language-specific summarization algorithms. Then, it presents the most recent advances in multi- and cross-lingual text summarization. The chapter classifies the presented methodology, highlights the main pros and cons, and discusses the perspectives of the extension of the current research towards cross-lingual summarization systems.

DOI: 10.4018/978-1-5225-9373-7.ch001

INTRODUCTION

In recent years, accomplice the recent advances of Web-based applications, the number of textual documents produced and made available in electronic form has steadily increased. To peruse potentially large collections of textual documents, domain experts often need for the aid of automatic compression tools, namely the document summarizers. These systems are able to produce informative yet succinct summaries by filtering out irrelevant or redundant content and by selecting the most salient parts of the text.

Text summarization is an established branch of research, whose main goal is to study and develop summarization tools which are able to extract high-quality information from large document collections (Tan et al., 2006). Plenty of approaches to document summarization have been proposed in literature. They commonly rely on Natural Language Processing (NLP), Information Retrieval (IR), or text mining techniques (Nazari & Mahdavi, 2019). Automated summarization systems have found application in industrial and research domains, e.g., content curation for medical applications (Zitnik et al., 2019), news recommendation (Tang et al., 2009), disaster management (Li et al., 2010), and learning analytics (Cagliero et al., 2019, Baralis & Cagliero, 2018).

The text summarization process commonly entails the following steps:

1. Filter the content of the input documents and transform it using ad hoc textual data representations.
2. Identify the key concepts mentioned in the text and extract significant descriptions of these concepts in textual form.
3. Generate summaries of the original document content that cover all of the salient concepts with minimal redundancy.

Statistics- and semantics-based text analyses are commonly applied in order to detect the most significant concepts and their descriptions in the text (Conroy et al., 2004). Most of them rely on the hypothesis that the content of all the original documents is written in the same language. This simplifies both the models used to capture in the text, which are usually language- and domain-specific, and the computation of text similarity measures, which usually rely on frequency-based term analyses. Hereafter, they will denote as "language-specific" summarizers all the systems that cannot be applied to documents written in different languages.

The rapid growth of Internet worldwide has produced a huge mass of textual documents written in a variety of different languages. Accessing the information contained in documents written in different languages has become a relevant yet compelling research issue (Wang et al., 2018). For instance, the findings described