Chapter 9

NewSum: “N-Gram Graph”–Based Summarization in the Real World

George Giannakopoulos
NCSR “Demokritos”, Greece & SciFY Not-for-Profit Company, Greece

George Kiomourtzis
SciFY Not-for-Profit Company, Greece & NCSR “Demokritos”, Greece

Vangelis Karkaletsis
NCSR “Demokritos”, Greece

ABSTRACT

This chapter describes a real, multi-document, multilingual news summarization application, named NewSum, the research problems behind it, as well as the novel methods proposed and tested to solve these problems. The system uses the representation of n-gram graphs in a novel manner to perform sentence selection and redundancy removal for the summaries and faces problems related to topic and subtopic detection (via clustering) and multi-lingual applicability, which are caused by the nature of the real-world news summarization sources.

INTRODUCTION

Automatic summarization has been under research since the late 50’s (Luhn, 1958) and has tackled a variety of interesting real-world problems. The problems faced range from news summarization (Barzilay & McKeown, 2005; Huang, Wan, & Xiao, 2013; Kabadjov, Atkinson, Steinberger, Steinberger, & Goot, 2010; D. Radev, Otterbacher, Winkel, & Blair-Goldensohn, 2005; Wu & Liu, 2003) to scientific summarization (Baralis & Fiori, 2010; Teufel & Moens, 2002; Yeloglu, Milios, & Zincir-Heywood, 2011) and meeting summarization (Erol, Lee, Hull, Center, & Menlo Park, 2003; Niekrasz, Purver, Dowding, & Peters, 2005).

The significant increase in the rate of content creation due to the Internet and its social media aspect, moved automatic summarization research to a multi-document requirement, taking into account the redundancy of information
across sources (Afantenos, Doura, Kapellou, & Karkaletsis, 2004; Barzilay & McKeown, 2005; J. M Conroy, Schlesinger, & Stewart, 2005; Erkan & Radev, 2004; Farzindar & Lapalme, 2003). Recently, the fact that the content generated by people around the world is clearly multilingual, has urged research to revisiting summarization under a multilingual prism (Evans, Klavans, & McKeown, 2004; Giannakopoulos et al., 2011; Saggion, 2006; Turchi, Steinberger, Kabadjov, & Steinberger, 2010; Wan, Jia, Huang, & Xiao, 2011).

However, this volume of summarization research does not appear to have reached a wider audience, possibly based on the evaluated performance of automatic systems, which consistently perform worse than humans (John M Conroy & Dang, 2008; Hoa Trang Dang & Owczarzak, 2009; Giannakopoulos et al., 2011).

In this chapter, we show how a novel, multilingual multi-document news summarization method, without the need for training, can be used as an everyday tool. We show how we designed and implemented an automatic summarization solution, named NewSum, which summarizes news from a variety of sources, using language-agnostic methods. We describe the requirements studied during the design and implementation of NewSum, how these requirements were met and how people evaluated the outcome of the effort.

Our main contributions in this chapter are, thus, as follows:

- We briefly study the requirements of a real-world summarization application, named NewSum. We describe task-aware specifications based on user and application context limitations (e.g. device, communication), source limitations and legal limitations.
- We describe a generic, language-agnostic method for extractive summarization, taking into account redundancy constraints. The method needs no training and minimizes the effort of crossing language boundaries, since it functions at the character level.
- We describe an open architecture for responsive summarization on a mobile setting.
- We provide an evaluation of the system based on non-expert evaluations, to represent market applicability of the system.

In the following section we provide some background on automatic summarization to sketch the related summarization research.

**BACKGROUND**

In this section, we briefly discuss summarization methods and systems that have been available as either research efforts, but also as real applications. We refer to the projects that aim at summarization and sketch the current state-of-the-art of the summarization sub-domains of salience detection and redundancy removal.

Summarization has been defined as a reductive transformation of a given set of texts, usually described as a three-step process: selection of salient portions of text, aggregation of the information for various selected portions, (optionally) abstraction of this information and, finally, presentation of the final summary text (S. Jones, 1999; I. M. Mani & Bloedorn, 1999). The summarization research community addresses major problems that arise during the summarization process.

- How can one group texts into topics, given a big set of texts of varying topics?
- How can one detect and select salient information to be included in the summary (ideally without training)?
- How can one avoid redundant or repeated information in the output summary, especially when multiple documents are used as input to the summarization process?