Chapter 9
Scaling and Semantically-Enriching Language-Agnostic Summarization

George Giannakopoulos
https://orcid.org/0000-0003-2459-589X
NCSR Demokritos, Greece & SciFY PNPC, Greece

George Kiomourtzis
SciFY PNPC, Greece & NCSR Demokritos, Greece

Nikiforos Pittaras
NCSR Demokritos, Greece & National and Kapodistrian University of Athens, Greece

Vangelis Karkaletsis
NCSR Demokritos, Greece

ABSTRACT

This chapter describes the evolution of a real, multi-document, multilingual news summarization methodology and application, named NewSum, the research problems behind it, as well as the steps taken to solve these problems. The system uses the representation of n-gram graphs to perform sentence selection and redundancy removal towards summary generation. In addition, it tackles problems related to topic and subtopic detection (via clustering), demonstrates multi-lingual applicability, and—through recent advances—scalability to big data. Furthermore, recent developments over the algorithm allow it to utilize semantic information to better identify and outline events, so as to offer an overall improvement over the base approach.

DOI: 10.4018/978-1-5225-9373-7.ch009

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). More recently, document summarization has moved on to specific genres and domains, such as (micro-)review summarization (Nguyen, Lauw & Tsaparas, 2015; Gerani, Carenini & Ng, 2019) and financial summarization (Isonuma et al, 2017).

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). We should note at this point, however, that even summary evaluation itself is a challenging scientific topic (Lloret, Aker & Plaza, 2018).

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