Chapter 116
Approaches to Large-Scale User Opinion Summarization for the Web

William Darling
Xerox Research Centre Europe, France

ABSTRACT

This chapter discusses approaches to applying text summarization research to the real-world problem of opinion summarization of user comments. Following a brief overview of the history of research in text summarization, the authors consider large scale user opinion summarization on the Web, a summarization problem that is distinct from the traditional domain that the research has focused on until very recently. More specifically, they consider opinion summarization of large datasets that generally include large degrees of noise and little editorial structure. To deal with this kind of real-world problem, the chapter addresses three major areas that must be considered and adhered to when designing systems for this type of problem: simple techniques, domain knowledge, and evaluative testing. Each area is covered in detail, and throughout the chapter, the lessons are applied to a case study that aims to apply the recommendations to designing a real-world opinion summarization system for a fictional book publisher.

INTRODUCTION

The research literature in automatic text summarization is extensive and spans nearly seven decades. The focus in this research has traditionally centered most strongly on summarization of news articles (Gong & Liu, 2001; Nenkova, et al., 2006; Haghighi & Vanderwende, 2009; Nenkova & McKeown, 2011), but in recent years has expanded in diverse directions to, most notably, opinion summarization (Hu & Liu, 2004; Blair-Goldensohn, et al., 2008; Lerman, et al., 2009; Lu, et al., 2009). Opinion summarization is of immense interest to corporations and governments, who can act more effectively when they understand the viewpoints of their customers or citizens that are being laid bare each day in Internet forums and social networks, but are so disparate and numerous that they cannot be efficiently digested without some service in between the raw data and the ultimate consumer.

DOI: 10.4018/978-1-5225-1759-7.ch116

Part of the impetus of the disparate evolution in text summarization's domain can be attributed to its application to real-world problems. While news summarization is interesting, in practice it is often unnecessary due to the concise and structured nature of newswire writing and the common existence of an already-included high quality succinct summary: the headline. Further, news summarization (as far as the task has commonly been defined in the literature) is arguably not a "real-world problem"; the datasets are small, the payoff is low, and the task is generally accomplished easily by humans.

In contrast, the areas that summarization research is expanding to are primarily those that can be considered real-world problems. Here, the payoffs are high, the datasets are often huge (making the problems interesting by virtue of fitting into the *in vogue* area of "big data'") and the tasks are those that, rather than simplifying or optimizing the efficiency of an existing practice, could not be done without the help of algorithmic – often machine learning based – approaches. The principal example is opinion summarization (Liu & Zhang, 2012). This broad field encompasses disparate research tasks that include, *inter alia*, topic modeling, sentence scoring, clustering, sentiment analysis, subjectivity prediction, and user modeling. Building a powerful and mature system for real-world use in this area necessitates expertise in at least a large subset of all of these areas, plus the experience and knowledge required to make them work together synoptically.

What it does not necessarily require, however, is "advanced" or overly complicated models. Peter Norvig and others have argued convincingly that in general more data with simple models beats less data with complex models (Halevy, et al., 2009). While Norvig's statement was with respect to machine learning in general, and in particular to cases where large amounts of data are available, the sentiment at least appears to be particularly apt with respect to text summarization. Models that have existed for several years tend to perform on par with more recent complicated methods and the simpler models are often easier to implement and scale much better. When these simple methods are coupled with domain-specific targeted assumptions about the structure of the input and the types of meta information that are available, impressive results can be easily achieved. In this Chapter, we explain powerful yet simple techniques for opinion summarization that build on these ideas.

This Chapter gives a practical overview to large-scale user opinion summarization particularly through the web. It broadly addresses three major areas that must be considered and adhered to when designing modern opinion summarization systems that tackle the real-world problems associated with understanding a collection of individual's viewpoints from opinionated text data. Here, we consider real-world problems to consist of tasks where there are large amounts of unstructured data, there is generally no gold-standard or notion of an objectively correct answer, and that can in some way be considered as *practical* or *commercial* in that there is a conceivable situation where the system would be commercially viable (in either an industrial or governmental context). While each of the three major areas necessarily feed on each other, they will be presented sequentially. They include: (1) relying on simple techniques; (2) domain knowledge and specialization; and (3) testing and evaluation. Following discussions of how best to utilize existing techniques and how to design and undertake competent testing and evaluation of a real-world user opinion summarization system, the Chapter will conclude with a brief look at where summarization research is headed in the near future. Throughout the Chapter, the recommendations will be exemplified by application to an informal case study involving the implementation of a book review summarization system for a publishing company. The case study considers specific existing techniques to implement, the domain knowledge that might be included, a proper evaluation framework, and how to tie all of the parts together into a coherent whole.