Chapter 7
Summarization in the Financial and Regulatory Domain

Jochen L. Leidner
https://orcid.org/0000-0002-1219-4696
Refinitiv Labs, UK & University of Sheffield, UK

ABSTRACT

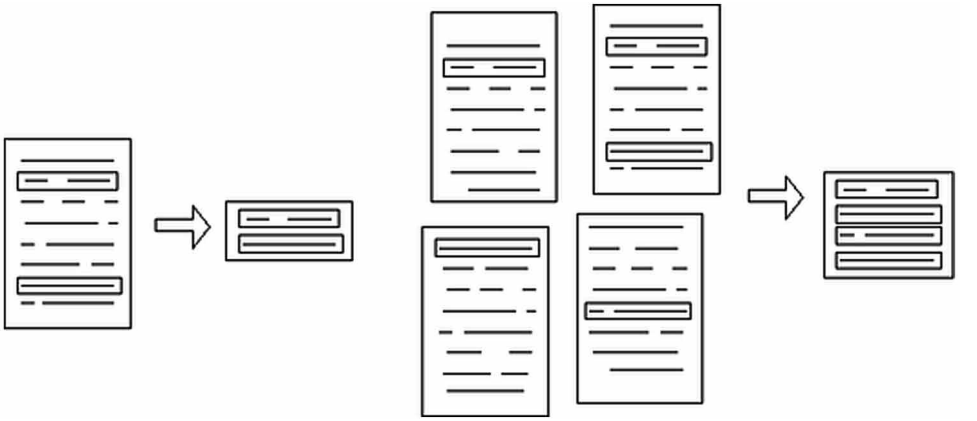
This chapter presents an introduction to automatic summarization techniques with special consideration of the financial and regulatory domains. It aims to provide an entry point to the field for readers interested in natural language processing (NLP) who are experts in the finance and/or regulatory domain, or to NLP researchers who would like to learn more about financial and regulatory applications. After introducing some core summarization concepts and the two domains are considered, some key methods and systems are described. Evaluation and quality concerns are also summarized. To conclude, some pointers for future reading are provided.

INTRODUCTION

Inderjeet Mani defined the goal of *automatic summarization* (also "summarisation" in British English) as "*to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's need"* (Mani, 2001). Therefore, the business value of it lies in its potential for enhancing the productivity of human information consumption (Modaresi *et al.*, 2017): the output of the task of summarizing an input text document comprising English prose is a shorter new document or shorter version of the original document that conveys most of the most important information

DOI: 10.4018/978-1-5225-9373-7.ch007

Figure 1. Single-document summarization (left) versus multi-document summarization (right).



contained in the original document, yet takes less time to read than the original full document.

Traditionally, we can distinguish between *single document summarization*, which takes as input a single document (*source document*) that needs to be summarized, and *multi-document summarization*, which takes as input a set of documents covering the same topic or topic area (Figure 1). In both cases, a single document, the summary (*target document*) is to be created. We can further distinguish between *extractive summarization*, which computes summaries by selecting text spans (phrases, sentences, passages) from the original document or documents, and *abstractive summarization*, which extracts pieces of information in a pre-processing step, and then constructs a synthetic new document, which is a summary that communicates said extracted facts, or it may even introduce new language not found in the source document(s) (Figure 2, right). Mathematically speaking, extractive summarization can be seen as a sequence of projections. Extractive summarization have the advantage of circumventing the problem of how to generate grammatical sentences as it merely selects from existing sentences; it has the disadvantages that a sequence of selected sentences may not make for smooth reading, as it is hard to combine them so as to maintain cohesion (broadly, to be linked together well at the micro-level) and coherence (roughly, to form a meaningful and logical text at the macro-level). The history of automatic summarization goes back to the German researcher Hans Peter Luhn, who worked on automatic summarization at IBM, where he created the method for extractive single-document summarization now named after him (Luhn, 1958).[1]

We can also distinguish between various kinds of methods. *Heuristic methods* like the Luhn method (outlined below) typically use a human-conceived scoring