Chapter 9

Graph-Based Abstractive Summarization: Compression of Semantic Graphs

Balaji Jagan
Anna University, India

Ranjani Parthasarathi
Anna University, India

Geetha T. V.
Anna University, India

ABSTRACT

Customization of information from web documents is an immense job that involves mainly the shortening of original texts. Extractive methods use surface level and statistical features for the selection of important sentences. In contrast, abstractive methods need a formal semantic representation, where the selection of important components and the rephrasing of the selected components are carried out using the semantic features associated with the words as well as the context. In this paper, we propose a semi-supervised bootstrapping approach for the identification of important components for abstractive summarization. The input to the proposed approach is a fully connected semantic graph of a document, where the semantic graphs are constructed for sentences, which are then connected by synonym concepts and co-refering entities to form a complete semantic graph. The direction of the traversal of nodes is determined by a modified spreading activation algorithm, where the importance of the nodes and edges are decided, based on the node and its connected edges under consideration.

1. INTRODUCTION

Text Summarization can be classified as extractive and abstractive methods. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document to produce a compressed form of the original text. The importance of the sentences is decided based on the statistical and linguistic features of sentences. In contrast, an abstractive summarization method consists
Graph-Based Abstractive Summarization

of understanding the original text and rephrasing it into different forms without changing the meaning conveyed in the original text, but in a compressed form of a summary. When compared with an extractive summary, the abstractive summary is a difficult and challenging task, which requires the semantic representation of the text, inference rules and natural language generation (Erkan & Radev 2004).

Extraction involves concatenating extracts taken from the corpus into a summary, whereas abstraction involves generating novel sentences from information extracted from the corpus. It has been observed that in the context of multi-document summarization of news articles, extraction may be inappropriate because it may produce summaries which are overly verbose or biased towards some sources (Barzilay et al., 1999). Extractive summarization (Gupta & Lehal 2010) includes selecting important information, paragraphs etc. from a document and combining it to form a new paragraph called as summary. The choice of the sentences depends upon statistical and linguistic features of the sentences. Extractive summaries are formulated by weighting the sentences as a function of high frequency words. Here, the most frequently occurring or the most favourably positioned text is considered to be the most important.

Abstractive summarization (Khan & Salim 2014) includes understanding the main concepts and relevant information of the main text and then expressing that information in short and clear format. Abstractive summarization techniques can again be classified into two categories- structured based and semantic based methods. Structured based approaches determines the most important information through documents by using templates, extraction rules and other structures such as tree, ontology etc. Semantic based approaches determines the most important information through, conceptual graphs, semantic networks, semantic graphs, etc. Abstractive summarization methods produce more coherent, less redundant and information rich summary. Generating abstract using abstractive summarization methods is a difficult task since it requires more semantic and linguistic analysis.

In general, the text summarization task is performed at various levels, such as the surface, entity and discourse (Hahn & Mani 2000). Surface-level approaches tend to represent information in terms of shallow parsers which can then be selectively combined to yield a selection function used to extract important information. Entity-level approaches (Mani & Maybury 1999) build an internal representation of the text, modeling text entities and their relationships. Text entities are units of texts, such as words, phrases, sentences or even paragraphs. These approaches tend to represent patterns of connectivity in the text to help determine what is salient. Discourse-level approaches (Mann & Thompson 1988) model the structure of the text and its relation to communicate goals.

Summarization is also carried out using graph-based approaches, such as LexRank (Erkan & Radev 2004) and TextRank (Mihalcea & Tarau 2004). LexRank has been applied to multi-document summarization, whereas TextRank has been applied to single document summarization and keyword extraction. Both the approaches apply a random walk in a fully connected undirected graph, to redistribute the node weights where text units (i.e. sentences) are represented as nodes and the similarities between the text units are represented as edges.

Most existing approaches rely on surface information where the important sentences are selected, and are simply added to the summary without understanding the meaning conveyed by the sentences. The disadvantages of these approaches result in summaries with redundant sentences, or less relevant or irrelevant sentences which are not important to the summary. To overcome these difficulties, a semantic interpretation of documents is required. Semantic information can be word level or context level. Word level semantics represents the semantic class associated with the word, such as concepts and attributes. On the other hand, context level semantics represents the semantic relationships existing between two different concepts. One such representation, where both word level and context level semantics are