Chapter 7
Keyword Extraction Based on Selectivity and Generalized Selectivity

Slobodan Beliga
University of Rijeka, Croatia

Ana Meštrović
University of Rijeka, Croatia

Sanda Martinčić-Ipšić
University of Rijeka, Croatia

ABSTRACT

This chapter presents a novel Selectivity-Based Keyword Extraction (SBKE) method, which extracts keywords from the source text represented as a network. The node selectivity value is calculated from a weighted network as the average weight distributed on the links of a single node and is used in the procedure of keyword candidate ranking and extraction. The selectivity slightly outperforms an extraction based on the standard centrality measures. Therefore, the selectivity and its modification – generalized selectivity as the node centrality measures are included in the SBKE method. Selectivity-based extraction does not require linguistic knowledge as it is derived purely from statistical and structural information of the network and it can be easily ported to new languages and used in a multilingual scenario. The true potential of the proposed SBKE method is in its generality, portability and low computation costs, which positions it as a strong candidate for preparing collections which lack human annotations for keyword extraction.

INTRODUCTION

The task of keyword extraction (KE) is to automatically identify a set of terms that best describe the document (Mihalcea & Tarau, 2004). Automatic keyword extraction establishes a foundation for various natural language processing applications: information retrieval, the automatic indexing and classifica-

DOI: 10.4018/978-1-5225-5042-6.ch007

tion of documents, automatic summarization and high-level semantic description (Balaji, Geetha, & Parthasarathi, 2016; Brian & Pradeep, 2010; Cheng & Qu, 2009), etc.

Although the keyword extraction applications usually work on single documents (document-oriented task) (Boudin, 2013; Lahiri, Choudhury, & Caragea, 2014; Palshikar, 2007), keyword extraction is also applicable to a more demanding task, i.e. the keyword extraction from a whole collection of documents (Dostal & Jezek, 2011; Grineva, Grinev, & Lizorkin, 2009; Jones & Paynter, 2002) (collection-oriented task) or from an entire web site (Wu & Agogino, 2003). In the era of big-data, obtaining an effective method for automatic keyword extraction from huge amounts of multi-topic textual sources is a nowadays necessity.

State-of-the-art keyword extraction approaches are based on statistical, linguistic or machine learning methods (Siddiqi & Sharan, 2015; Beliga, Meštrović, & Martinčić-Ipšić, 2015). In the last decade the focus of research has shifted towards unsupervised methods, mainly towards network or graph enabled keyword extraction. In a network enabled keyword extraction the document representation may vary from very simple (words are nodes and their co-occurrence is represented with links), or can incorporate very sophisticated linguistic knowledge like syntactic (Lahiri et al., 2014; Liu & Hu, 2008; Mihalcea & Tarau, 2004) or semantic relations (Grineva et al., 2009; Joorabchi & Mahdi, 2013; Wang, Wang, Senzhang, & Zhoujun, 2014; Bougouin, Boudin, & Daille, 2016; Martinez-Romo, Araujo, & Duque Fernandez, 2016; Rafiei-Asl & Nickabadi, 2017; Ying et al., 2017).

Typically, the source (document, text, data) for keyword extraction is modeled with one network. This way, both the statistical properties (frequencies) as well as the structure of the source text are represented by a unique formal representation, hence a complex network.

A network enabled keyword extraction exploits different measures for the task of identifying and ranking the most representative features of the source – the keywords. The keyword extraction powered by network measures can be on the node, network or subnetwork level (Beliga et al., 2015). Measures on the node level are: degree, strength (Lahiri et al., 2014); on the network level: coreness, clustering coefficient, PageRank motivated ranking score or HITS motivated hub and authority score (Boudin, 2013; Mihalcea & Tarau, 2004); on the subnetwork level: communities (Rafiei-Asl & Nickabadi, 2017; Grineva et al., 2009). Most of the research was motivated by various centrality measures: degree, betweenness, closeness and eigenvector centrality (Ludwig, Thiel, & Nürnberger, 2017; Abilhoa & de Castro, 2014; Lahiri et al., 2014; Boudin, 2013; Matsuo, Ohsawa, & Ishizuka, 2001; Palshikar, 2007; Mihalcea & Tarau, 2004).

This research presents the novel selectivity-based method for the unsupervised keyword extraction from the co-occurrence network of texts. A new network measure – the node selectivity, originally proposed by Masucci and Rodgers (2006, 2009) (that can distinguish an original network from a shuffled one), is applied to automatic keyword extraction. Selectivity is defined as the average weight distributed on the links incident to the single node. Furthermore, we utilize a generalized selectivity measure defined according to the generalized weighted degree originally proposed by Opsahl, Agneessens and Skvoretz (2010). In previous work, the node selectivity measure performed in favor of the differentiation between original and shuffled Croatian texts (Margan, Meštrović, & Martinčić-Ipšić, 2014a; Margan, Martinčić-Ipšić, & Meštrović, 2014b), for the differentiation of text genres (Martinčić-Ipšić, Miličić, & Meštrović, 2016) and for prediction of missing links in networks (Martinčić-Ipšić, Močibob, & Perc, 2017).

The node selectivity measure has been preliminary tested for the keyword extraction task in our early work (Beliga, Meštrović, & Martinčić-Ipšić, 2014), where we explore the potential of the selectivity measure for the keyword extraction in Croatian news articles. The full potential of selectivity was ana-