Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art

Kazi Saidul Hasan and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
{saidul,vince}@hlt.utdallas.edu

Abstract

State-of-the-art approaches for unsupervised keyphrase extraction are typically evaluated on a single dataset with a single parameter setting. Consequently, it is unclear how effective these approaches are on a new dataset from a different domain, and how sensitive they are to changes in parameter settings. To gain a better understanding of state-of-the-art unsupervised keyphrase extraction algorithms, we conduct a systematic evaluation and analysis of these algorithms on a variety of standard evaluation datasets.

1 Introduction

The keyphrases for a given document refer to a group of phrases that *represent* the document. Although we often come across texts from different domains such as scientific papers, news articles and blogs, which are labeled with keyphrases by the authors, a large portion of the Web content remains untagged. While keyphrases are excellent means for providing a concise summary of a document, recent research results have suggested that the task of automatically identifying keyphrases from a document is by no means trivial. Researchers have explored both supervised and unsupervised techniques to address the problem of automatic keyphrase extraction. Supervised methods typically recast this problem as a binary classification task, where a model is trained on annotated data to determine whether a given phrase is a keyphrase or not (e.g., Frank et al. (1999), Turney (2000; 2003), Hulth (2003), Medelyan et al. (2009)). A disadvantage of supervised approaches is that they require a lot of training data and yet show bias towards the domain on which they are trained, undermining their ability to generalize well to new domains. Unsupervised approaches could be a viable alternative in this regard.

The unsupervised approaches for keyphrase extraction proposed so far have involved a number of techniques, including language modeling (e.g., Tomokiyo and Hurst (2003)), graph-based ranking (e.g., Zha (2002), Mihalcea and Tarau (2004), Wan et al. (2007), Wan and Xiao (2008), Liu et al. (2009a)), and clustering (e.g., Matsuo and Ishizuka (2004), Liu et al. (2009b)). While these methods have been shown to work well on a particular domain of text such as short paper abstracts and news articles, their effectiveness and portability across different domains have remained an unexplored issue. Worse still, each of them is based on a set of assumptions, which may only hold for the dataset on which they are evaluated.

Consequently, *we have little understanding of how effective the state-of the-art systems would be on a completely new dataset from a different domain*. A few questions arise naturally. How would these systems perform on a different dataset with their original configuration? What could be the underlying reasons in case they perform poorly? Is there any system that can generalize fairly well across various domains?

We seek to gain a better understanding of the state of the art in unsupervised keyphrase extraction by examining the aforementioned questions. More specifically, we compare five unsupervised keyphrase extraction algorithms on four corpora with varying domains and statistical characteristics. These algorithms represent the ma-

Coling 2010: Poster Volume, pages 365–373,
Beijing, August 2010