Machine Learning-based Extraction of Keyphrases and its Applications in Multiple Domains

Gábor Berend

Advisors:
Dr. János Csirik
Dr. Richárd Farkas

University of Szeged
PhD School in Computer Science

Summary of PhD Thesis

Szeged, 2014

Motivation

The pace at which data is generated nowadays has encouraged the introduction and application of intelligent and automated ways to process data. Data items today are generated by an extremely wide audience and range – including geolocational, acceleration measurement data – mostly due to smart devices. Despite the abundant appearance of novel types of data, it is still extensively produced in the traditional textual format. These textual data items may take many forms, ranging from (micro)blog and forum posts to news portal entries and scientific literature. Documents originating from the different genres can differ greatly – in their length, writing style, degree of structure in them, and so on. However, what they all have in common is that large quantities of them are accessible and that their detailed processing is practically impossible without machine-augmented methods.

Knowing the most important phrases of textual documents can provide a condensed representation for them, which can make their processing easier. However, the manual determination of the sets of important phrases for every single document in a large collection of documents is a tedious and expensive task and it often requires expert knowledge. **Natural language processing** techniques – mostly relying on **machine learning** – can fortunately help the automatic generation of keyphrases for documents.

In this thesis, various models for the extraction of keyphrases from textual documents of various genres and languages are presented, and their potential end-application utilization is demonstrated in the form of a document visualization system. Although most of the earlier studies focused on the domain of scientific papers, we will introduce models for the extraction of keyphrases in two languages (i.e. English and Hungarian) and from various genres including scientific publications, news articles and product reviews as well.

Structure of the dissertation

The results of this thesis can be divided into two logical parts. The first part deals with the generation of keyphrases from textual documents of various genres and languages and the second part illustrates how the outputs of these models can be utilized in applications.

In Chapter 3, we introduce the problem of retrieving keyphrases from documents originating from news articles, a genre being more heterogeneous from both topical and stylistic perspectives compared to scientific publications – which is the standard domain for performing keyphrase extraction. Another special point of this chapter is that it describes the extraction of keyphrases from documents written in Hungarian.

In Chapter 4, we describe a machine learning-based feature rich keyphrase extraction framework for the standard setting of keyphrase extraction from English scientific documents. The novel features presented in this part of the thesis are designed to capture the semantic orientation of keyphrases by means of linguistic analysis and external knowledge sources.

In Chapter 5, we introduce the task of extracting keyphrases from another, quite dissimilar genre, namely opinionated utterances from product review sites. Here, we verify our hypothesis that *pro* and *con* phrases assigned to product reviews behave in a similar way to keyphrases of non-opinionated documents. Furthermore, we propose useful extensions to general keyphrase extraction models to achieve better performance measures on this specialized, opinion mining-related keyphrase extraction task.

In Chapter 6, we focus on the utilization of the outputs of the keyphrase extraction models described in earlier chapters. The possible applications described here are a technique for assigning sets of keyphrases to document subcorpora (in contrast to single documents) and a keyphrase-based corpus visualization approach.

Venue	Year		Chapter 3	4	5	6
SEMEVAL	2010	[3]	•			
RANLP	2011	[9]	•			
NLE	2014	under review	•			
ECAI	2010	[8]		•		
IJCNLP	2011	[1]			•	
RANLP	2011	[7]			•	
WASSA	2012	[6]			•	
CICLing	2013	[5]				•
IJCNLP	2013	[4]				•

Table 1: The relation between the thesis topics and the corresponding publications

In Table 1, we list the author's key publications related to this thesis. For a full list of the author's publications, please visit `http://www.inf.u-szeged.hu/~berendg/?pp=publ`.

Summary of the thesis results

The main goal of this thesis was to demonstrate techniques for the generation of keyphrases extracted from textual documents of various types, i.e. news articles, scientific documents and product reviews. The proposed solutions take into account the special aspects of the domains and rely on extra-textual world knowledge by utilizing Wikipedia. Here, we briefly summarize the main results of the thesis.

Keyphrase Generation from Newswire

In Chapter 3, we presented our system specially constructed for the assignment of keyphrases for the news articles comprising the archive of the news portal Origo. This task was special in that we not only had to handle the morphological

richness of the Hungarian language, but also ensure that the keyphrases assigned to the news articles behave coherently at the level of the entire document collection (e.g. by making sure that they did not contain synonymous expressions). Another special aspect of the task was that we had to take special care with the so-called *abstract keyphrases*, i.e. keyphrases that are not otherwise present in the document for which they need to be assigned. Combining our approaches into a framework, we managed to achieve a document-level precision of 75.44%, which was well beyond the prior expectations of the employees of Origo. Related to his publication [8], the author regards the following as his main contributions to the research topic:

- Introduction of a ranking procedure for selecting the most likely keyphrases
- Assignment of abstract keyphrases to documents based on definitions derived from Wikipedia
- Various ways for the assignment of abstract keyphrases to documents based on the link structure of Wikipedia

Keyphrase Extraction from Scientific Documents

In Chapter 4, we proposed novel ways of exploiting extra-textual information during the extraction of keyphrases. Besides these features being useful in the domain of scientific publications – as Chapter 5 verified it – they tend to be successfully applicable for different domains as well, implying their wide-range applicability. Based on evaluations carried out on multiple datasets, our proposed method performs competitively with state-of-the-art systems. One of the main advantages of the proposed method is that even though it relies on Wikipedia like some other approaches, it does not require a full index to be created from all the textual contents of Wikipedia, as it relies only on its category structure. Although our approach requires fewer resources, it can still perform competitively with

other methods. Related to his publications [2, 3, 9], the author regards the following as his main contributions to the research topic:

- Extension of the existing keyphrase candidate filtering techniques
- Introduction of a condensed representations for sequential features
- Utilization of extra-textual information for representing keyphrase candidates

Opinion Phrase Extraction

In Chapter 5, we verified our hypothesis that keyphrase extraction techniques can be employed to the task of determining important aspects of product reviews, i.e. *pro and con* expressions. Besides pointing out the applicability of standard keyphrase extraction approaches for this task, we suggested several domain-specific features. Incorporating these features into standard keyphrase extraction frameworks resulted in significant gains in performance. In our experiments, we also investigated the subjective nature of judging the appropriateness of keyphrases. Domain differences among product reviews were demonstrated via cross-product experiments within the domain of product reviews. It was also shown how the severe drop in the quality of the extracted keyphrases in such cases can be lessened by using domain adaptation methods. Related to his publications [1, 7], the author regards the following as his main contributions to the research topic:

- Extension of standard keyphrase extraction to opinion phrase extraction
- Creation of a manually annotated opinion phrase corpus
- Verification of the applicability of domain adaptation techniques in opinion phrase extraction

Applications of keyphrase extraction

In Chapter 6, we showed how keyphrases could be utilized in a visualization framework. The chapter proposed a solution for the assignment of keyphrases for subcorpora on the basis of information theoretic considerations and the sets of keyphrases determined for the individual documents. Moreover, it was also shown in that chapter how topics can be inferred from document-level keyphrases and then be visualized based on the overlap between the pairs of documents in a text collection. Related to his publications [4, 5], the author regards the following as his main contributions to the research topic:

- Proposing a method for assigning keyphrases to document subcorpora
- Applying single-document keyphrases-based document representation
- Introducing the keyphrase similarity graph-based clustering and visualization framework

Bibliography

[1] Gábor Berend. Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/I11-1130`.

[2] Gábor Berend. Exploiting extra-textual information in keyphrase extraction. *Natural Language Engineering*, page to appear, 2014.

[3] Gábor Berend and Richárd Farkas. SZTERGAK: Feature engineering for keyphrase extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 186–189, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S10-1040`.

[4] Gábor Berend and Richárd Farkas. Keyphrase-driven document visualization tool. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, pages 17–20, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/I13-2005`.

[5] Gábor Berend and Richárd Farkas. Single-document key-

phrase extraction for multi-document keyphrase extraction. *Computación y Sistemas*, 17(2):179–186, 2013.

[6] Gábor Berend and Veronika Vincze. How to evaluate opinionated keyphrase extraction? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 99–103. Association for Computational Linguistics, 2012.

[7] Gábor Berend, István T. Nagy, György Móra, and Veronika Vincze. Inter-domain opinion phrase extraction based on feature augmentation. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 41–47, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. URL `http://aclweb.org/anthology/R11-2006`.

[8] Richárd Farkas, Gábor Berend, István Hegedűs, András Kárpáti, and Balázs Krich. Automatic free-text-tagging of online news archives. In *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 529–534, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press. ISBN 978-1-60750-605-8. URL `http://dl.acm.org/citation.cfm?id=1860967.1861071`.

[9] István Nagy T., Gábor Berend, and Veronika Vincze. Noun compound and named entity recognition and their usability in keyphrase extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 162–169, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. URL `http://aclweb.org/anthology/R11-1023`.