Network-Enabled Keyword Extraction for Under-Resourced Languages

Slobodan Beliga and Sanda Martinčić-Ipšić

Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51 000 Rijeka, Croatia
{sbeliga, smarti}@inf.uniri.hr

Abstract. In this paper we discuss advantages of network-enabled keyword extraction from texts in under-resourced languages. Network-enabled methods are shortly introduced, while focus of the paper is placed on discussion of difficulties that methods must overcome when dealing with content in under-resourced languages (mainly exhibit as a lack of natural language processing resources: corpora and tools). Additionally, the paper discusses how to circumvent the lack of NLP tools with network-enabled method such is SBKE method.

Keywords: Network-enabled keyword extraction · Under-resourced languages · NLP tools · SBKE method

1 Introduction

Automatic keyword extraction is the process of identifying key terms, phrases, segments or words from a textual content that can appropriately represent the main topic of the document [1, 14]. Keyword extraction (KE) methods can be roughly divided into three categories: supervised, semi-supervised and unsupervised [1]. Network-enabled or graph-based are considered as unsupervised KE methods.

Today the automatic keyword extraction from texts still remains an open question, especially for content written in under-resourced languages. For under-resourced languages there are no reliable tools which can be used for keyword extraction task and text preprocessing, such as: POS and MSD taggers, stemmers, lemmatisers, stop-words lists, lexical resources like WordNet, controlled vocabularies, benchmark or monitoring datasets, and other tools or resources.

The main aim of this work is to discuss the problems of keyword extraction in under-resourced languages and as the possible solution we recommend network or graph-enabled KE methods. These methods use knowledge incorporated in the structure of network or graph to extract keywords and therefore circumvent unavailable linguistic tools required in a certain KE method development.

In the second part of this paper we will explain the concept of under-resourced languages and describe the problems that occur in KE methods for such languages. The third part of the paper will explain the general procedure of network-enabled KE methods, more precisely, SBKE method through the lenses of portability to different languages. Moreover, we provide a list of available benchmark datasets for KE development and evaluation in order to illustrate the problem of the lack of resources. The paper ends with some concluding remarks and presentation of plans for future work.

2 Deficiencies of KE Methods for Under-Resourced Languages

Next we explain the concept of under-resourced languages in the context of text analysis and keyword extraction task, and then we describe the problems that occur in KE methods for some of the European languages which have been considered as under-resourced.

2.1 Under-Resourced Languages

Today there are more than 6900 languages in the world and only a small fraction of them is supported with the resources required for implementation of Natural Language Processing (NLP) technologies or applications [2]. Authors in [2] explained that main stream NLP is mostly concerned with languages for which large resources are available or which have suddenly become of concern because of the economic interest or political influence.

The term “under-resourced languages” was introduced by Krauwer (2003) and complemented by Berment (2004). They both define criteria to consider a particular language as under-resourced: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists, etc. [3, 4]. Other authors have used the terms “low-density” or “less-resourced” instead of “under-resourced” languages. Further, Berment in [4] categorizes human languages into three categories, based on their digital “readiness” or presence in cyberspace and software tools: “tau”-languages: totally-resourced languages, “mu”-languages: medium-resourced languages and “pi”-languages: under-resourced languages [4]. In addition to individual researchers, these issues are recognized as important for group of researchers, and commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders gathered in Multilingual Europe Technology Alliance (META). META network is dedicated to fostering the technological foundations of a multilingual European information society with a vision of Europe united as one single digital market and information space for Language Technology [8]. In META White Paper Series the state of language technology development is categorized into the following areas: Machine Translation, Speech Processing, Text Analysis, and Speech and Text Resources. Within these areas languages can be classified into following categories: excellent, good, moderate, fragmentary and weak/no support. The most important area for KE is Text Analysis in which the languages marked with ‘+’ in Table 1 have the lowest support [9]. Since the META is European alliance, data presented in Table 1 are related exclusively with European languages, as well as the scope of this paper.

It is important to notice that the list of systematized languages in Table 1 may not be an exhaustive list of European under-resourced languages for the area of text analysis. However, there may be additional under-resourced languages such as Bosnian or Albanian which are not listed because no relevant studies were reported.

Besides to languages that are in weak support category, there are other languages that are classified into fragmentary category and few of them in moderate. As expected, English is the only language with good support in all areas (see Table 2). Expressed in the proportions: weak supported - 30%, fragmentary supported - 50%, moderate supported - 16.66% and good supported - 3.33%.

Table 1. Cross-language comparison of European languages classified according to the areas into weak/no support category [9].

Language	Machine Translation	Speech Processing	Text Analysis	Speech and Text Resources
Bulgarian	+			
Croatian	+	+	+	
Czech	+			
Danish	+			
Estonian	+		+	
Finnish	+			
Greek	+			
Icelandic	+	+	+	+
Irish	+		+	+
Latvian	+	+	+	+
Lithuanian	+	+	+	+
Maltese	+	+	+	+
Portuguese	+			
Serbian	+		+	
Slovak	+			
Slovene	+			
Swedish	+			
Welsh	+	+	+	+

Table 2. Cross-language comparison of European good, moderate, and fragmentary supported languages in Text Analysis area [9].

Language	Good	Moderate	Fragmentary
English	+		
Dutch		+	
French		+	
German		+	
Italian		+	
Spanish		+	
Basque			+
Bulgarian			+

Catalan			+
Czech			+
Danish			+
Finnish			+
Galician			+
Greek			+
Hungarian			+
Norwegian			+
Polish			+
Portuguese			+
Romanian			+
Slovak			+
Slovene			+

Besides META systematization, credibility and objectivity of belonging to under-resourced category are also measured with BLARK (Basic Language Resource Kit) concept. BLARK is defined as the minimal set of language resources that is necessary to do any precompetitive research and education at all [3]. It must be under 10 out of 20 in order to be considered as under-resourced language. A BLARK comprises criteria, such as: written language corpora, spoken language corpora, mono and bilingual dictionaries, terminology collections, grammars, annotation standards and tools, corpus exploration and exploitation tools, different modules (e.g. taggers, morphological analyzers, parsers, speech recognizers, text-to-speech), etc. [3].

2.2 Problems in Keyword Extraction and Motivation

Information Retrieval (IR) and Natural Language Processing (NLP) experts which set their research focuses on keyword extraction task, at the ACL workshop on novel computational approaches to keyphrase extraction from 2015, detected several open problems [15]: technical term extraction using measures of neology, decompounding for keyphrase extraction (especially for German language – compound morphology), extracting social oriented keyphrase semantics from Twitter, applications to noun compounds syntax and semantic, problem of over-generation errors in automatic keyword or keyphrase extraction, which is also known problem in network-enabled methods.

Another important issue but rarely discussed in the context of KE is a lack of tools for KE method development for under-resourced languages. Although there are numerous keyword extraction methods for richer-resourced languages with remarkable performance such as methods presented in [7, 10, 11, 12] (both in supervised or unsupervised setup), in the absence of language tools it is difficult to adopt them for other languages, especially for under-resourced languages. These methods are most often developed for the English language. In other words, language scalability (portability) of these methods is limited to a particular language or language group. In order to support multilingualism, and circumvent poor portability, we propose unsupervised methods, graph- or network-enabled methods for keyword extraction. Network structure enables representation of the input text as graph or network, regardless of language. In a network representation of the input text the nodes (vertices) are unique words and the edges (links) between two nodes are established when two words share a relation (e.g. co-occur within a window).

An example of graph-based method is Selectivity-Based Keyword Extraction (SBKE) [14]. Instead of developing new tools for a language of interest, application of this method requires only tuning of various parameters which are inherent for particular language (fine tuning of parameters for candidate extraction, setting the filtering thresholds for keyword expansion, …).

3 Network-Enabled KE Concept

In a network approach, network of words is used for the representation of texts, which enables the exploration of the relationships and structural information incorporated in text very efficiently. Although there are different variations, the most common way of document modeling into graph is the representation where words are modeled by vertices (nodes) and their relations are represented by edges (links). The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a corpus. On this basis there are various possibilities for the analysis of a network structure (topology) and we will focus on the most common – network structure of the linguistic elements themselves using co-occurrence relations. This is a basic relation, but it has shown effective results in numerous studies, such as in [5, 6, 7]. Another reason to use co-occurrence, and not any semantic or syntactic relation is the lack of language tools which could extract these relations.

Fig. 1. Co-occurrence network constructed from text: *"KEYSTONE - semantic keyword-based search on structured data sources" is a COST Action aiming to make it straightforward to search through structured data sources like databases using the keyword-based search familiar to many internet users. The scientific objective of KEYSTONE is to analyse, design, develop and evaluate techniques to enable keyword-based search over large amounts of structured data."*

Figure 2. presents the generalized process for portability of network-enabled keyword extraction techniques. In the first step keyword candidates are extracted from the text. After that, candidates are filtered according to properties specific for particular method. Note that in this step various network measures can be used for rankings: closeness, degree or betweenness centrality, TextRank, etc. In the final step, candidates are ranked according to the obtained value from the used measure and used thresholds, resulting with a candidate list of keywords.

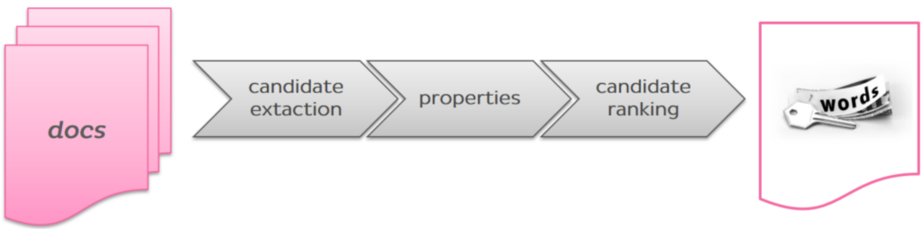

Fig. 2. Generalization of the keyword extraction techniques.

3.1 SBKE Method Portability for Under-Resourced Languages

SBKE – Selectivity-Based Keyword Extraction method is a network-enabled method for keyword extraction which consists of two phases: **(1) keyword extraction** and **(2) keyword expansion**. The node selectivity value is calculated from the weighted network as the average weight distributed on the links of a single node and is then used in the procedure of keyword candidate ranking and extraction [13, 14]. This method does not require linguistic knowledge as it is derived purely from statistical and structural information of the network, therefore it is suitable for many European under-resourced languages. The main advantage is that networks are constructed from pure co-occurrence of words in the input texts. Moreover, the method achieves results which are above the TF-IDF (Term Frequency - Inverse Document Frequency) baseline for English and Croatian language [14].

As previously mentioned, SBKE method consist of two phases decomposed into several steps. First phase: **(1) keyword extraction**: in initial step, it is advisable that the text is preprocessed: lemmatized or stemmed (depends on tools availability for stemming or lemming in particular language). Although, preprocessing is not necessary because SBKE works without stemming or lemmatization, but it is advisable to preprocess the input text in order to reduce the size of the network, which is of importance in highly inflectional languages. After that, language network can be constructed from preprocessed input using the co-occurrence of words. For constructed network the selectivity or generalized selectivity of each node is measured as indicated in [14]. Additionally, parameters of generalized selectivity can be tuned individually for particular language or corpus. In the second phase: **(2) keyword expansion**, keyword candidates are expanded to longer sequences – two or three words long keyphrases (according to the weight of links with neighboring nodes in the network). Sequence construction is derived solely from the properties of the network. In other words, the method does not require any intensive language resources or tools except light preprocessing. However, preprocessing can be omitted as well. Finally, the method is portable to under-resourced languages because it does not require linguistic knowledge as it is derived purely from statistical and structural information of the network.

3.2 Textual Resources for KE in Under-Resourced Languages

If we want to compare the performance of the automatic KE with humans, then a valid method for the evaluation of the KE method can be carried out only by bench-mark datasets which contain keywords annotated by human experts. Some of the available datasets are presented in Table 3. It shows only those data sets that are annotated by humans (usually students involved in individual studies or human experts in a particular area). Most of the available datasets are in the English language, while other available datasets cover the French (DEFT – scientific articles published in social science journals [16]); French and Spanish (FAO 780 – FAO publications with Agrovoc terms [25, 30]); Polish (abstracts of academic papers for PKE method [17]), Portuguese (tweet dataset [5] and news transcriptions [32]), and Croatian (HINA – news articles [18]). Other languages, especially under-resourced languages which are in our focus do not have developed datasets for keyword extraction task. Collection of comparable Lithuanian, Latvian and Estonian laws and legislations (available in [19]) could be used for facilitated dataset development for KE task. However, it would be necessary to invest into human experts' annotations of keywords for the evaluation purposes.

Table 3. Available datasets with annotated keywords by human per language, number of annotators, size in the number of documents and usage of controlled vocabulary. Controlled Vocabulary is marked with yes/no when controlled vocabulary was assumed, but not always obeyed.

Lang.	Dataset	Controlled Vocabulary	Annotat.	Num. of documents	Description
ENGLISH	SemEval2010 [22]	yes/no	-authors,-readers,-authors and readers combined	trial: 40 training: 144 testing: 100	Student annotators from the Computer Science department of the National University of Singapore.
	Wiki20 [25]	yes (Wikipedia)	15 teams	20	Computer Science papers, each annotated with at

					least 5 Wikipedia articles by 15 teams of indexers.
	CiteULike [25]	no	330 volunteers	180	Publications crawled from CiteULike, keywords assigned by different CiteULike users who saved these publications.
	FAO 30 [25, 30]	yes (thesaurus)	6 experts	30	Food and Agriculture Organization (FAO) of the United Nations publications.
	500N-KPCrowd [31]	yes	20 HITs	500 (450+50)	only the key phrases selected by at least 90% of the annotators
	Krapivin [29]	-	author assigned and editor corrected keyphrases.	2000	Scientific papers from computer science domain published by ACM.
	Wan and Xiao [28]	-	-author -students	308	Documents from DUC2010, including ACM Digital Library, IEEE Xplore, Inspec and PubMed articles, author-assigned keyphrases and occasionally reader-assigned
	Nguyen and Kan [27]	-	-one by original author -one or more by student annotators	120	Computer science articles, author-assigned and reader assigned keyphrases undergraduate CS students.
	INSPEC [26]	yes two sets of keywords (Inspec thesaurus) no	professional annotator	2000 training: 1000 validation: 500 testing: 500	Abstracts of journal articles present in Inspec, from disciplines Computers and Control, and Information Technology. Both the controlled terms and the uncontrolled terms may or may not be present in the abstracts.
	Twitter dataset [23, 24]	-	11 humans	1827 tweets training: 1000 development:	The annotations of each annotator are combined by

				327 testing: 500	selecting keywords that are chosen by at least 3 annotators.
	Email dataset [21]	-	2 annotators	349 emails: 225 threads: 124	Email dataset consists of single and thread emails.
ENGLISH FRENCH SPANISH	FAO 780 [25, 30]	yes (Agrovoc thesaurus)	-human annotator	-780 English -60 French -47 Spanish indexers working independently.	FAO publications with Agrovoc terms. Documents are indexed by one indexer.
POLISH	PKE [17]	yes/no	1 expert (author of the paper)	12000 training: 9000 testing: 3000	Abstracts from Polish academic papers downloaded from web sources (e.g. pubmed, yadda). All abstracts have at least 3 keywords.
FRENCH	DEFT [16]	yes (50%)	author	234 training: 60% testing: 40%	French scientific articles published in social science journal.
		no (50%)	students	234 training: 60% testing: 40%	
CROATIAN	HINA[18]	yes/no	8 human experts	1020 training: 960 testing: 60	Croatian news articles from the Croatian News Agency (HINA).
PORTUGUESE	Portuguese tweet dataset TKG method [5]	no	3 users	300 tweets	Portuguese tweet collections from 3 Brazilian TV shows: ‘Trofeu Imprensa’, 'A Fazenda' and 'Crianca Esperanca'.
	110-PT-BN-KP Marujo [32]	-	-one annotator	110 news training: 100 testing: 10	The gold standard is made of 8 BN programs - 110 news subset (transcriptions), from the European Portuguese ALERT BN database.

Datasets with controlled vocabulary consist of manually annotated keywords by humans using only words from original text, titles of Wikipedia articles or some predefined list of allowed words as the controlled vocabulary. Such datasets are particularly suitable for methods which are not able to generate new words. Human annotators are also an important determinant of KE dataset - the quality of the dataset is higher if the number of human (individuals or teams) annotators is higher. Having only a single set of keywords assigned by a human annotator (individual or collaborating team) per document, taking it as the gold standard, and using the popular measures of precession, recall and their harmonic mean, F1, to evaluate the quality of keyword assigned by the automatic machine annotator ignores the highly subjective nature of key-word annotation tasks [20]. In this case Inter-Indexer Consistency (IIC) can be used instead. IIC measures the quality of keywords assigned to the test documents by developed method with those assigned by each team or human annotators.

3.3 Preliminary Results

In the absence of datasets for KE in under-resourced languages (with keywords annotated by human experts or another machine algorithm), it is not possible to evaluate the SBKE method in standard measures (recall, precision, F-measure or IIC-Inter-Indexer Consistency). However, we show some preliminary results for the Serbian language. All extracted keywords from Serbian news articles available on the web portal www.novosti.rs from 3 different genres: politics, economics and sports are listed in the Table 4. It seems that SBKE method for the Serbian language prefers open-class words (such as nouns, adjectives, etc.), that are good candidates for real keywords. This was also the case for Croatian [13], and expected, since they are related Slavic languages.

Table 4. Keywords extracted from 3 different texts written on Serbian language from political, economic and sports genres.

Genre	Title	Keywords (translated to English)
POLITICS	Migrants	refugees, political, life, Angela Merkel, elections, united, more, Austria, options, Germany, population, year
ECONOMICS	Credit without a permanent job	customers, credit, capable, banks, ability, loan, institutions, interest, rates, reserve, categories, evaluation, mandatory, contract, criteria, agent
SPORTS	Serbian paralympic athletes traveled to the Rio	athlete, Rio, support, champion, medal, pride, minister, preparation, effort, table, tennis, team, London, Uroš Zeković

4 Conclusion

This paper briefly describes graph or network-enabled keyword extraction methods. It also explains why these methods are suitable for under-resourced languages. We provide the detailed list of datasets for keyword extraction for EU languages. Using graph-based methods for keyword extraction can open the possibilities for the development of other applications which in its initial phase require keywords.

In future work we will try SBKE method for other under-resourced languages to show that knowledge incorporated in the network should replace non-existing linguistic tools necessary for keyword extraction from semi-structured web sources. In particular, we will focus on KE dataset development for Serbian, Estonian, Latvian, Lithuanian, Maltese and possibly other non-European under-resourced languages.

References

1. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: An Overview of Graph-Based Keyword Extraction Methods and Approaches, Journal of Information and Organizational Sciences, 39(1), 1-20 (2015)
2. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic Speech Recognition for Under-resourced Languages: A Survey. Speech Communication, vol. 56, 85-100, (2014)
3. Krauwer, S.: The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In Proceedings of the 2003 International Workshop Speech and Computer SPECOM-2003, pp. 8-15, Moscow, Russia (2003)
4. Berment, V.: Méthodes pour informatiser des langues et des groupes de langues "peu dotées". PhD Thesis, J. Fourier University – Grenoble I, (2004)
5. Abilhoa, W. D., & Castro, L. N.: A keyword extraction method from twitter messages represented as graphs. Applied Mathematics and Computation, 240, 308-325 (2014)
6. Palshikar, G. K.: Keyword Extraction from a Single Document Using Centrality Measures. Pattern Recognition and Machine Intelligence (LNCS) Second International Conference, PReMI 2007, Springer Berlin Heidelberg, vol. 4815, pp. 503-510 (2007)
7. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. In Proceedings of Empirical Methods in Natural Language Processing – EMNLP 2004. Barcelona, Spain: ACL, pp. 404-411 (2004)
8. META-NET – official site (May 2016), http://www.meta-net.eu/
9. META-NET White Paper Series: Key Results and Cross-Language Comparison (May 2016), http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison
10. Joorabchi, A., Mahdi, A. E.: Automatic keyphrase annotation of scientific documents using Wikipedia and genetic algorithms. Journal of Information Science, 39(3), 410–426 (2013)
11. Lahiri, S., Choudhury, S.R., Caragea, C.: Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks. arXiv preprint arXiv:1401.6571, (2014)
12. Grineva, M., Grinev, M., Lizorkin, D.: Extracting key terms from noisy and multitheme documents. ACM 18th conference on World Wide Web, pp.661–670, (2009)
13. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Toward Selectivity-Based Keyword Extraction for Croatian News". CEUR Proceedings of the Workshop on Surfacing the Deep and the Social Web (SDSW 2014), Vol. 1310, pp. 1-8, Riva del Garda, Trentino, Italy (2014)
14. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Selectivity-Based Keyword Extraction Method. International Journal on Semantic Web and Information Systems (IJSWIS), 12(3), 1-26 (2016)
15. Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction, ACL-IJCNLP 2015, Beijing, China (2015)
16. Paroubek, P., Zweigenbaum, P., Forest, D., Grouin, C.: Indexation libreet controlee d'articles scientifiques. Presentation et resultats du defi fouille de textes DEFT2012. In Proceedings of the DEfi Fouille de Textes 2012 Workshop, pp. 1–13, (2012)
17. Kozłowski, M.: PKE: a novel Polish keywords extraction method. Pomiary Automatyka Kontrola, R.60(5), 305-308 (2014)
18. Mijić, J., Dalbelo-Bašić, B., Šnajder, J.: Robust Keyphrase Extraction for a Large-Scale Croatian News Production System. In Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Balkan Languages, Zagreb, Croatia: Croatian Language Technologies Society, pp. 59-66 (2010)

19. Collection of comparable Lithuanian, Latvian and Estonian laws and legislations (June 2016), http://metashare.nlp.ipipan.waw.pl/metashare/repository/browse/collection-of-comparable-lithuanian-latvian-and-estonian-laws-and-legislations/8d0d633eae7711e2a28e525400c0c5ef33b6cfc6ca074e1ab58859157c8374e7/#
20. Zunde, P., Dexter M. E.: Indexing consistency and quality. American Documentation, 20(3), 259–267 (1969)
21. Loza, V., Lahiri, S., Mihalcea, R., Lai, P.: Building a Dataset for Summarization and Keyword Extraction from Emails. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 2441-2446, Reykjavik, Iceland, (2014)
22. Su, N. K., Medelyan, O., Min-Yen, K., Timothy, B.: Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation. 47(3), 723-742 (2013)
23. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., et al.: Part-of-speech tagging for twitter: annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, HLT '11, Stroudsburg, PA, USA. Association for Computational Linguistics (2011)
24. Marujo, L., Wang, L., Trancoso, I., Dyer, C., Black, A. W., Gershman, A., et al.: Automatic Keyword Extraction on Twitter. ACL (2015).
25. Medelyan, O. Human-competitive automatic topic indexing. PhD thesis. Department of Computer Science, University of Waikato, New Zealand (2009)
26. Hulth, A.: Improved Automatic Keyword Extraction Given More Linguistic Knowledge. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 216-223 (2003)
27. Nguyen, T. D., Kan, M.: Key phrase Extraction in Scientific Publications. Proceeding of International Conference on Asian Digital Libraries, pp. 317-326 (2007)
28. Wan, X., Xiao, J.: CollabRank: towards a collaborative approach to single-document keyphrase extraction. In Proceedings of COLING, pp. 969–976 (2008)
29. Krapivin, M., Autaeu, A., Marchese, M.: Large dataset for keyphrase extraction. Technical Report DISI-09-055, DISI, University of Trento, Italy (2009)
30. Medelyan, O., Witten, I. H.: Domain independent automatic keyphrase indexing with small training sets. Journal of American Society for Information Science and Technology. Vol. 59 (7), pp. 1026-1040 (2008)
31. Marujo, L., Gershman, A. Carbonell, J., Frederking, R., Neto, J. P.: Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization. Proceedings of LREC 2012 (2012)
32. Marujo, L., Viveiros, M., Neto, J. P.: Keyphrase Cloud Generation of Broadcast News. In proceeding of 12th Annual Conference of the International Speech Communication Association, Interspeech (2011)