Keyword Extraction from Parallel Abstracts of Scientific Publications

Slobodan Beliga[1](✉), Olivera Kitanović[2], Ranka Stanković[2], and Sanda Martinčić-Ipšić[1]

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

[2] Faculty of Mining and Geology, University of Belgrade,
Đušina 7, 11 000 Belgrade, Serbia
{olivera.kitanovic,ranka.stankovic}@rgf.bg.ac.rs

Abstract. In this paper, we study the keyword extraction from parallel abstracts of scientific publication in the Serbian and English languages. The keywords are extracted by a selectivity-based keyword extraction method. The method is based on the structural and statistical properties of text represented as a complex network. The constructed parallel corpus of scientific abstracts with annotated keywords allows a better comparison of the performance of the method across languages since we have the controlled experimental environment and data. The achieved keyword extraction results measured with an F1 score are 49.57% for English and 46.73% for the Serbian language, if we disregard keywords that are not present in the abstracts. In case that we evaluate against the whole keyword set, the F1 scores are 40.08% and 45.71% respectively. This work shows that SBKE can be easily ported to new a language, domain and type of text in the sense of its structure. Still, there are drawbacks – the method can extract only the words that appear in the text.

Keywords: Graph-based keyword extraction
Bilingual keyword extraction · SBKE method · Parallel abstracts

1 Introduction

The task of keyword extraction is to automatically identify a set of terms that best describes the document [1,2]. Keyword extraction can be a demanding task, especially when the aim is keyword extraction from bilingual or multilingual textual sources. In such a case, a keyword extraction method should be insensitive to natural language or appropriate for extraction in different natural languages at the same time.

One of the open research questions in the keyword extraction task is to develop a method that is general enough for keyword extraction in several languages simultaneously. Therefore, we focus on a method that can be easily ported to a new language. The prerequisites needed for this desirable characteristic are

J. Szymański and Y. Velegrakis (Eds.): IKC 2017, LNCS 10546, pp. 44–55, 2018.
https://doi.org/10.1007/978-3-319-74497-1_5

that the method does not require deeper linguistic preprocessing. We believe that this is especially important for extraction tasks in low-resourced languages that have less developed language tools [10]. In other words, more sophisticated keyword extraction methods in the text preprocessing step usually use some heuristics to gain in performance by using semantic or syntactic knowledge. As the source of syntactic knowledge, methods usually use part-of-speech tags (POS) in order to restrict access to certain types of words (e.g. nouns, verbs or adjectives) [14,19,20] or suffix sequences which denote the sequence of morphological suffixes of its words [27,29].

Wikipedia is one of the most commonly used semantic sources: using n-grams that appear in Wikipedia article titles as candidates for keywords [22], utilizing Wikipedia as a thesaurus for candidate selection from documents' content [21], exploiting links on Wikipedia to detect keywords candidates [24] or using terminological databases to encode the salience of candidate keyphrases [28]. The methods can be also based on extracted noun-phrase chunks that satisfy predefined lexico-semantic patterns [23]. These different approaches for keyword extraction are effective on various textual sources, such as scientific articles [26], news articles [8,9], blogs [22], meeting transcripts [20], emails [25], web pages [27], etc. However, if such a keyword extraction method needs to be applicable in a bilingual or multilingual environment then the module which incorporates semantic or linguistic knowledge needs to be developed for each language separately.

In this paper, we test the applicability of a graph-enabled method called the selectivity-based keyword extraction method (SBKE) proposed in [8] for the bilingual keyword extraction task. The dataset consists of parallel Serbian-English abstracts from scientific articles from the domain of geology and mining including annotated keywords by the authors of articles. In the scientific literature, methods were studied and compared in different languages: besides the most studied – English language [2,8,13,19,21,22,25,26] are Portuguese [3], Polish [7], Croatian [8], French and Spanish [4–6]. However, no studies report the extraction from parallel texts of different languages with bilingual keyword annotations. To the best of our knowledge, this will present a graph-based keyword extraction for parallel abstracts in scientific articles for the first time, as well as a new bilingual keyword extraction dataset. The main contributions of this paper are:

(1) the development of bilingual (Serbian-English) keyword extraction dataset, and
(2) the comparative study of the effectiveness of the SBKE method for keyword extraction on parallel texts written in two languages.

In addition, in this work we test whether the SBKE method is portable to a new language, in this research Serbian (in addition to Croatian and English [8]), to a new domain of geology and mining (in addition to domains of news and technical reports from Wikipedia [8]) and finally, that SBKE can be applicable to short texts, hence abstracts from scientific articles.

In Sect. 2, we provide a description of the methodology. First, we briefly explain the SBKE keyword extraction method (Subsect. 2.1), then we provide a brief overview of the used NLP tools for Serbian and English (Subsect. 2.2) and used evaluation methodology (Subsect. 2.3). The description of the used parallel dataset for English and Serbian languages is in Sect. 3. Section 4 presents the results, while the concluding remarks and future plans are in Sect. 5.

2 Methodology

A detailed description of the selectivity-based keyword extraction method (SBKE) is available in [8]. However, in the following section we will explain the basic characteristics of the method to ensure the readability and completeness of the manuscript.

2.1 The SBKE Method

The network or graph-based approach, where a network (or graph) of words is used for the representation of texts, enables the exploration of the relationships and structural information incorporated in a text very efficiently. Although, there are variations, the usual way of representing documents as a graph models words as vertices (nodes) and their relations as edges (links). The weight of the link is proportional to the overall co-ccurrence frequencies of the corresponding word pairs within a corpus. We will focus on the network construction around co-occurrence relations of adjacent words within sentences, since it requires no semantic or syntactic preprocessing of the input text. Network enabled keyword extraction methods exploit various structural properties (usually centrality measures) of the nodes in a network for the extracting and ranking of keyword candidates [1].

The 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 [8,9]. The node in/outselectivity and generalized in/out-selectivity values are calculated from a directed weighted network as the average weight distributed on the ingoing/outgoing links of the single node and used in the procedure of keyword candidate ranking and extraction. This method does not require linguistic knowledge (apart from stemming or lemmatization) as it is derived purely from the statistical and structural information of the network [10].

In this study, we use the SBKE method on a directed and weighted network. An individual network is constructed separately for each Serbian and for each English text. More preciously, from all the constructed networks, we rank the nodes according to the highest in/out-selectivity values above a threshold greater than 1, as proposed in [8]. Therefore, we obtain two sets of extracted keywords, one for the Serbian, and one for the English version of the text. Preserving the same threshold value in all documents resulted in a different number of extracted nodes (one-word long keyword candidates) from each network, which is the union of the highly-ranked nodes according to the in/out-selectivity values for the particular language.

2.2 Text Preprocessing Tools

Serbian is a highly inflectional Slavic language. Although we use the keyword extraction method designed with light or no linguistic knowledge, some text preprocessing is needed and includes the conversion of the input text to lowercase, the removal of misspelled symbols and lemmatization. In a similar way, we preprocessed the English text: converted to lowercase and stemmed using the Porter stemmer. Stemming and lemmatization are also needed for a better matching of the extracted and annotated keywords during evaluation to overcome differences between the inflected forms in the text and the lemmatized keyword forms of the same word.

In the text preprocessing stage for the English language we use:

(1) Stop-word list - extracted from the Natural Language Toolkit (NLTK) for Python [11], and
(2) the Porter stemmer [11] for stemming as a procedure to map all words with the same stem to a common form (stem). Its main use is as part of the term normalization process (removing the inflectional suffixes from words).

For preprocessing of texts in the Serbian language we use:

(1) Stop-word list - prepared at the Human Language Technology Group at the University of Belgrade [30], and
(2) a Serbian lemmatizer. For lemmatization, we use Serbian morphological electronic dictionaries and grammars developed within the University of Belgrade Human Language Technology Group [17]. Morphological electronic dictionaries of Serbian for NLP have been developing for many years now. In the dictionary of lemmas (DELAS) each lemma is described in full detail so that the dictionary of forms containing all the necessary grammatical information (DELAF) can be generated from it, and subsequently used for various NLP tasks. Serbian e-dictionaries of simple forms have reached a considerable size: they have more than 140,000 lemmas generating more than 5 million forms and 18,000 multi-word lemmas [18].

Different approaches (stemming and lemmatization) were caused by the differences in morphological feature of these two languages. The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Lemmatization is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or a dictionary form. Stemming usually refers to a process that chops off the ends of words in the hope of achieving this goal correctly most of the time. Serbian, like other Slavic languages, is a highly-inflected language, with complex grammatical rules that cannot be adequately expressed by stemming rules. However, for highly-inflected languages, lemmatization can hardly be avoided as each keyword can have many inflected forms (for multiword units from five to ten or even more). On the other hand, for English several efficient and accurate stemmers are available and we used the Porter stemmer, as one of the widely-used stemmers for the text preprocessing of the English language.

Both stemming and lemmatization play very important roles when it comes to increasing the relevance and recall capabilities of a retrieval system. When these techniques are used, the number of indexes used is reduced because the system is using one index to present several similar words that have the same root or stem [12].

2.3 Evaluation Methodology

The dataset used in this experiment contains only one set of annotated keywords – provided by the author(s) of each abstract (scientific paper). In the case when only one set of annotated keywords is available, the evaluation of the keyword extraction is performed as in the standard information retrieval tasks. Hence, precision (P), recall (R) and the $F1$ score are used for the evaluation. When comparing the performance of an automatic method (algorithm) with a human annotation, precision is calculated as the number of keywords in the intersection of a set of keywords annotated by a human (A) and a set of keywords annotated using algorithm (B) divided by the number of keywords annotated using the algorithm:

$$P = \frac{A \cap B}{B}. \tag{1}$$

The recall is calculated as the number of keywords in the intersection of a set of keywords annotated by a human and a set of keywords annotated using the algorithm divided by the number of keywords annotated by a human:

$$R = \frac{A \cap B}{A}. \tag{2}$$

The $F1$ score is the harmonic mean of precision and recall, calculated as:

$$F1 = \frac{2PR}{P + R}. \tag{3}$$

3 Textual Resources

In this experiment, we have used abstracts from the scientific journal "Underground Mining Engineering" published by the University of Belgrade, Faculty of Mining and Geology. Apart from technical underground mining related topics, the journal publishes papers from other fields of mining, geology, and geosciences, as well as from other scientific and technical disciplines having a direct or indirect application in mining. During the period of 2004–2012, the journal published 55 papers bilingually, in Serbian and in English. These papers are available online as aligned parallel text in the Biblisha[1] digital library, as well as separate documents. The Biblisha digital library contains scientific publications from other journals

[1] http://jerteh.rs/biblisha/ListaDokumenata.aspx?JCID=2&lng=en.

and also contains project reports that are published in two languages – Serbian and English. All the documents are provided with the usual metadata (article's author(s), publication date, title, keywords, abstract etc.) and are aligned at the sentence level [15,16].

For the research presented in this paper, we used a collection of 50 bilingual documents with approximately 4,800 aligned sentences. Since papers were published bilingually, they were already available both in Serbian and English, where most of the papers were originally written in Serbian and then translated into English by professional translators. Texts have various lengths, in Serbian the texts contain from 34 to 259 words (on average 100) and in English from 44 to 286 words (on average 110). The statistics of the used English and Serbian parallel abstract are presented in Table 1.

All the documents are supplied with metadata and keywords, annotated by human experts – the authors of the articles. The number of annotated keywords ranges from 3 to 18 in the Serbian and from 3 to 15 in the English texts (the average in both is 7). Scientists usually define keywords in their lemmatized form, while in the Serbian texts (and rarely in English) they appear in many inflected forms, which are different from lemma. Bilingual Serbian-English KE dataset introduced in this paper is publicly available from http://langnet.uniri.hr/resources.html.

The previous research [14] for terminology extraction in the Serbian language used the rule-based method for multi-word term extraction that relies on lexical resources for modeling various syntactic structures of multi-word terms. It is applied in several domains, also among them is the corpus of Serbian texts from the geology and mining domain containing more than 600,000 simple word forms. Part of this approach was the automatic elimination of less probable candidates: extracted and lemmatized multiword terms are filtered to reject falsely offered lemmas and then they are ranked by introducing measures that combine the linguistic and statistical information (C-Value, T-Score, LLR, and Keyness). In previous research, all the texts were joined and the entire collection was treated as a single text, while for the research presented in this paper, the text processing and analysis is performed per each text document in the collection. SBKE method does not include calculation of C-Value, T-Score, LLR, and Keyness, it follows the procedure described in Subsect. 2.1.

Table 1 presents the descriptive statistics for 50 parallel abstracts in the Serbian and English language including the average value, the minimal and maximal number of words in rows for each category presented by columns. The first column is related to the numbers of words in the text, the KW count lists the number of keywords given by an author, while KW in the text shows how many keywords given by an author are actually present in the abstract. The difference between the KW count and KW in text values depicts the number of OOV (out-of-vocabulary) annotated keywords.

Table 1. The statistics for 50 parallel abstracts in Serbian and English language.

	Serbian			English		
	#words	#KW	#KWinText	#words	#KW	#KWinText
Average	100.6	6.64	5.38	110.48	6.72	5.5
Min	34	3	2	44	3	2
Max	259	18	13	286	15	12

4 Results

The results of the experiments are presented in terms of R, P and the $F1$ score in Table 2. The left part of Table 2 presents the evaluation performance of the SBKE method according to the set of annotated keywords – provided by the author(s) of the abstracts. The right part of the table presents the evaluation according to the set of annotated keywords without out-of-vocabulary (OOV) words. All OOV words are removed from the set of keywords for the evaluation. The results are shown for one-word long keywords even though the SBKE can extract keywords that contain two or three words.

When analyzing the results, it is important to consider the fact that the keywords are specified by domain scientists and this can be highly subjective, comprising of a lot of background knowledge on the topic. Sometimes their approach to keywords selection is oriented on the overall meaning and essence. Thereafter, in several cases the given keywords are not present in the text as the same term, in those cases, the concept is replaced with a synonym or hypernym.

For Serbian as well as for English languages, recall achieves higher values than precision (from 3% to 17%). This also holds for the Croatian and English languages as elaborated in our previous work [8] regardless of the inclusion or removal of OOV keywords, reflecting the greediness of the SBKE method. Note that the SBKE method is designed as "greedy" and extracts as much candidates as possible, which can cause the over-generation problem. Still, the SBKE method can circumvent the overgeneration problem by simple tuning of the filter applied to the weights during the expansion steps of setting the appropriate cut-off threshold during the extraction phase.

Note that the results in this study are better for the Serbian than for the English language (see Table 2). This is in line with our previous findings for the Croatian language [8]. Both, Serbian and Croatian language are morphologically rich, and closely related languages from South Slavic language family. Unlike English, which is inflectional language and has a strict word ordering in a sentence.

Next, in the right part of Table 2 (without OOV keywords) the evaluation results show that the SBKE method for Serbian achieves an F1 score of 49.57%, and for English an $F1$ score of 46.73%. The results of all measures (R, P, and $F1$) are generally higher when they are measured for keywords without OOV words. This is expected because people tagged ∼18% of the keywords in English, and ∼19% in Serbian that did not appear in the original texts.

So far, the SBKE method has been tested on longer English Wikipedia texts (with an average length of 5,919 words per text) and Croatian newspaper articles (with an average length of 335 words per text), where SBKE achieved F1 scores of 24.8% and 34.21% for Croatian and English respectively [8]. In the present study, the method is tested on abstracts of scientific articles with an average length of 100 words per abstract in Serbian and 110 per abstract in English – shorter texts. This is the reason why we stopped after the first phase (called keyword extraction in the SBKE method). Usually, SBKE performs better on longer texts (containing more information on the structural properties of the input text), but here we can explore the performance on the shorter texts [8]. The achieved results suggest that SBKE can be applied to shorter documents as well. Since SBKE is grounded in a structural and statistical information incorporated into the network structure the expected outcome is to achieve better performance on larger texts and on the whole document collection. In this case, SBKE proved correctly also on shorter texts. This outcome requires deeper further investigation which we plan to address in the future.

Moreover, the benefit of this work could be considered in future research, as the initial step in extracting concepts for the construction of the ontology for the domain of geology and mining in the Serbian language.

Finally, it is worth mentioning that the different structure and syntax of the Serbian and English languages are reflected in the results. By combining (translating) Serbian and English keywords, a larger set of keywords can be obtained. This is the advantage of bilingual keyword extraction, which standard methods for keyword extraction cannot reach, and remains an open question for future work.

Table 3 represents two different examples of abstracts in the test dataset in a form of a preprocessed texts. On the left side is an example where the SBKE method returned a larger set of keywords where the broader concept "method" is added to "analytic hierarchy processes (AHP)". On the right side of the table is an example where the word "speed" is listed as a keyword specified by a human, but in the text the author used a synonym term "velocity" (stemmed to: veloc). Since the word "speed" is not present in the original text, the method will never extract it as a keyword. Similar examples adversely affect the success of extraction and reduce the efficiency of the SBKE method in terms of $F1$ score. These examples

Table 2. Results of keyword extraction for parallel (Serbian-English) abstracts in scientific articles expressed in terms of R, P and an $F1$ score for all keywords defined by an author (in the left part) and for keywords without out-of-vocabulary (OOV) words (in the right part).

	All keywords			Without OOV		
	R [%]	P [%]	$F1$ [%]	R [%]	P [%]	$F1$ [%]
Serbian	54.32	45.96	45.71	63.38	45.96	49.57
English	44.62	41.20	40.08	55.58	44.48	46.73

Table 3. Two examples for extracted keywords compared to pre-assigned keywords on parallel Serbian-English abstracts. The keywords discussed in the example are underlined or written in *italic*.

Serbian	English
Author: analitički hijerarhijski proces (AHP)	**Author:** concret qualiti *speed* wave ultrasound
SBKE Method: *izbor* hijerarhijski analitički ahp proces *metod*	**SBKE Method:** concret qualiti wave ultrasound
Text: primena *metod* analitički hijerarhijski proces ahp kod *izbor* utovarni-transportni mašina. u ovaj rad prezentovan ona *metod* analitički hijerarhijski proces ahp i njen primena kod proces odlučivanje u rudarski inženjerstvo. konkretan u ovaj rad dat ona primena ahp *metod* kod *izbor* model utovarni-transportni mahati ne sa električni pogon na osnov utvrđen kriterijum odlučivanje kao i dodeljivanje težinski koeficijenata pojedin kriterijum, koji uticati na proces donošienje konačan odluka	**Text:** estim of the qualiti of built-in concret by the ultrasound observ. thi paper present result of the propag *veloc* investig of ultrasound wave in the concret construct so call non-destruct method in situ due to inspect of concret qualiti that is inbuilt into the bodi of the durutovici dam built for the pljevlja coal mine. by the *veloc* of ultrasound wave measur the follow paramet will be concret homogen presenc of gap crack and other defect in concret as well as concret qualiti relat to it strength

imply the possibility for the introduction of semantical knowledge into the further stages of the presented keyword extraction method. Namely, the list of extracted keyword candidates in the next stage can be expanded/corrected with semantic knowledge, with expansion to synonyms, hypernyms and/or hyponyms, which can be of high importance for application recall improvement.

5 Conclusion

In this work, we explored the keyword extraction from parallel abstracts of scientific papers from the domain of geology and mining in the Serbian and English languages. We show that the selectivity-based keyword extraction (SBKE) method is general enough to be easily ported to another language – Serbian, because it requires only shallow linguistic preprocessing. Then we tested the applicability of the SBKE method in a new and highly specialized scientific domain – a text collection from the geology and mining domain. Finally, the scientific abstract is limited to the number of characters, therefore we also test the applicability of SBKE on short texts.

The experimental part of the paper is focused on the performance of the SBKE method on parallel texts from the Serbian and English languages[2].

[2] Bilingual Serbian-English KE dataset is publicly available from http://langnet.uniri.hr/resources.html.

The new set-up of parallel texts enabled better insights into the performance across different languages simultaneously preserving the nature, size, and content of the texts. Usually, we compare unrelated datasets across languages. This set-up provides a controlled and fair environment for the evaluation.

We can conclude that SBKE can be easily ported to a different language, domain and type of text in the sense of its structure. Still, there are drawbacks, the method can extract only the words that appear in the text. However, we performed the evaluation with and without the out-of-vocabulary (OOV) keywords, showing that the results are promising even for the included OOV keywords.

In future work, we are planning to expand the keyword extraction from abstracts to whole scientific articles from the domain of mining and geology which are available in complete written form in both Serbian and English languages. It is important to compare keyword extraction results from whole papers with those extracted solely from short abstracts. Besides that, extracted keywords from whole papers can serve as a basis for the first approximation of a geological ontology construction. In addition, we will explore, if we can gain by translating the set of annotated keywords from the source to the target language and obtain larger sets of annotated keywords.

Acknowledgments. The authors would like to acknowledge networking support by the ICT COST Action IC1302 KEYSTONE – Semantic keyword-based search on structured data sources (www.keystone-cost.eu). The authors would also like to thank the University of Rijeka for the support under the LangNet project (13.13.2.2.07).

References

1. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: An overview of graph-based keyword extraction methods and approaches. J. Inf. Organ. Sci. **39**(1), 1–20 (2015)
2. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of Empirical Methods in Natural Language Processing - EMNLP 2004, pp. 404–411. ACL, Barcelona (2004)
3. 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)
4. Medelyan, O.: Human-competitive automatic topic indexing. Ph.D. thesis. Department of Computer Science, University of Waikato, New Zealand (2009)
5. Medelyan, O., Witten, I.H.: Domain independent automatic keyphrase indexing with small training sets. J. Am. Soc. Inf. Sci. Technol. **59**(7), 1026–1040 (2008)
6. Paroubek, P., Zweigenbaum, P., Forest, D., Grouin, C.: Indexation libre et 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)
7. Kozłowski, M.: PKE: a novel Polish keywords extraction method. Pomiary Autom. Kontrola **60**(5), 305–308 (2014)
8. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Selectivity-based keyword extraction method. Int. J. Sem. Web Inf. Syst. (IJSWIS) **12**(3), 1–26 (2016)

9. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Toward selectivity-based keyword extraction for croatian news. In: CEUR Proceedings of the Workshop on Surfacing the Deep and the Social Web (SDSW 2014), Riva del Garda, Trentino, Italy, vol. 1310, pp. 1–8 (2014)
10. Beliga, S., Martinčić-Ipšić, S.: Network-enabled keyword extraction for under-resourced languages. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) KEYSTONE 2016. LNCS, vol. 10151, pp. 124–135. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_11
11. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc., Sebastopol (2009)
12. Balakrishnan, V., Ethel, L.-Y.: Stemming and lemmatization: a comparison of retrieval performances. Lect. Notes Softw. Eng. **2**(3), 262–267 (2014)
13. Ludwig, P., Thiel, M., Nürnberger, A.: Unsupervised extraction of conceptual keyphrases from abstracts. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) KEYSTONE 2016. LNCS, vol. 10151, pp. 37–48. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_4
14. Stanković, R., Krstev, C., Obradović, I., Lazić, B., Trtovac, A.: Rule-based automatic multi-word term extraction and lemmatization. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia (2016). ISBN 978-2-9517408-9-1
15. Stanković, R., Krstev, C., Lazić, B., Vorkapić, D.: A bilingual digital library for academic and entrepreneurial knowledge management. In: Proceeding of 10th International Forum on Knowledge Asset Dynamics - IFKAD 2015: Culture, Innovation and Entrepreneurship: Connecting the Knowledge Dots, Bari, Italy, pp. 1764–1777 (2015)
16. Stanković, R., Krstev, C., Vitas, D., Vulović, N., Kitanović, O.: Keyword-based search on bilingual digital libraries. In: Calì, A., Gorgan, D., Ugarte, M. (eds.) KEYSTONE 2016. LNCS, vol. 10151, pp. 112–123. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53640-8_10
17. Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lazetić, G., Stanojević, M.: The Serbian Language in the Digital Age. META-NET White Paper Series. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30755-3. Rehm, G., Uszkoreit, H. (Series eds.)
18. Krstev, C., Vitas, D., Stanković, R.: A lexical approach to acronyms and their definitions. In: Mariani, Z.V.J. (ed.) Proceedings of the 7th Language & Technology Conference, pp. 219–223. Fundacja Uniwersytetu im. A. Mickiewicza, Poznan (2016)
19. Wan, X., Xiao, J.: Single document keyphrase extraction using neighborhood knowledge. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence, pp. 855–860 (2008)
20. Liu, F., Pennell, D., Liu, F., Liu, Y.: Unsupervised approaches for automatic keyword extraction using meeting transcripts. In: Proceedings of the HLT: The Annual Conference on Empirical Methods in NLP, pp. 257–266 (2009)
21. Joorabchi, A., Mahdi, A.E.: Automatic keyphrase annotation of scientific documents using Wikipedia and genetic algorithms. J. Inf. Sci. **39**(3), 410–426 (2013)
22. Grineva, M., Grinev, M., Lizorkin, D.: Extracting key terms from noisy and multi-theme documents. In: Proceedings of the 18th International Conference on World Wide Web, pp. 661–670. ACM, New York (2009)
23. Lahiri, S., Choudhury, S.R., Caragea, C.: Keyword and keyphrase extraction using centrality measures on collocation networks (2014). http://arxiv.org/pdf/1401.6571.pdf

24. Medelyan, O., Frank, E., Witten, I.H.: Human-competitive tagging using automatic keyphrase extraction. In: Proceedings of the 2004 Conference on Empirical Methods in NLP, pp. 1318–1327 (2009)
25. Lahiri, S., Mihalcea, R., Lai, P.-H.: Keyword extraction from emails. Nat. Lang. Eng. **23**(2), 295–317 (2016)
26. Kim, S.N., Medelyan, O., Kan, M.-Y., Baldwin, T.: SemEval-2010 task 5: automatic keyphrase extraction from scientific articles. In: SemEval 2010 Proceedings of the 5th International Workshop on Semantic Evaluation, Los Angeles, California, pp. 21–26 (2010)
27. Yih, W.-T., Goodman, J., Carvalho, V.R.: Finding advertising keywords on web pages. In: Proceedings of the 15th International Conference on World Wide Web, pp. 213–222 (2010)
28. Lopez, P., Romary, L.: HUMB: automatic key term extraction from scientific articles in GROBID. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 248–251 (2010)
29. Nguyen, T.D., Kan, M.-Y.: Keyphrase extraction in scientific publications. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 317–326. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77094-7_41
30. Utvic, M.: List of frequency corpus of contemporary Serbian language (in Serbian). In: Milanovic, A., Stanojcic, Ž., Popovic, Lj. (eds.) International Slavic Center, Faculty of Philology, vol. 43/3, pp. 241–262 (2014)