Single-Document Keyphrase Extraction for Multi-Document Keyphrase Extraction

Gábor Berend and Richárd Farkas

University of Szeged, Department of Informatics,
Árpád tér 2., 6720 Szeged,
Hungary

{berendg, rfarkas}@inf.u-szeged.hu

Abstract. Here, we address the task of assigning relevant terms to thematically and semantically related sub-corpora and achieve superior results compared to the baseline performance. Our results suggest that more reliable sets of keyphrases can be assigned to the semantically and thematically related subsets of some corpora if the automatically determined sets of keyphrases for the individual documents of an entire corpus are identified first. The sets of keyphrases assigned by our proposed method for the workshops present in the ACL Anthology Corpus over a 6-year period were considered better in more than 60% of the test cases compared to our baseline system when evaluated against an aggregation of different human judgements.

Keywords. Multi-document keyphrase extraction, knowledge management, information retrieval.

Extracción de palabras clave de documentos individuales para extracción de palabras clave de documentos múltiples

Resumen. En este artículo se considera el tema de asignación de términos relevantes a sub-corpus con temas y semántica relacionados y se logran resultados superiores a los del rendimiento de referencia. Los resultados obtenidos en este trabajo muestran que los conjuntos más confiables de palabras clave pueden ser asignados a subconjuntos con temas y semántica relacionados de un corpus si primero se identifican automáticamente los subconjuntos de palabras clave de documentos individuales en todo corpus. Los conjuntos de palabras clave asignados mediante el método propuesto para los talleres incluidos en ACL Anthology Corpus para el periodo de 6 años fueron considerados mejor en más de 60

Palabras clave. Extracción de palabras clave de documentos múltiples, administración de conocimiento, recuperación de información.

1 Introduction

The clustering and visualization of huge sets of documents is a widely used knowledge discovery technique. Assigning keyphrases to a cluster of documents makes the understanding of and navigation across these documents much easier for humans. In this paper, we propose a novel method for characterizing the main content of document sets with a few keyphrases.

We will focus on the more frequent and realistic scenario where keyphrases are not present by default for the documents comprising some corpora. The standard but still state-of-the-art approach for this task is based on the bag-of-words model and applies an information theoretical metric to rank the candidate phrases [8]. Here, we propose to pre-rank candidate phrases by using single-document keyphrase extraction techniques [14, 16]. The goal of these techniques is to represent the content of a single document with a few characteristic phrases. In contrast to the bag-of-words-based approach where the importance of the phrases of a document is solely represented by their number of occurrences (such as in the case of tf-idf weighting), positional and contextual information as well as information derived from external semantic knowledge resources are taken into account when selecting the candidate phrases.

For an empirical evaluation, we decided to assign keyphrases for the workshop papers of ACL Anthology and assessed how well they described the theme of a workshop. As the absolute human evaluation of keyphrases is highly subjective, we asked four researchers from the field of computational linguistics to compare the quality of the keyphrases of two systems with each other. The automatic evaluation of

keyphrases is usually based on string matching, which handles semantically (even closely) related phrases as a mismatch if their surface forms differ. We experimented with several automatic evaluation methods for the scientific domain and we shall introduce a procedure for evaluating the keyphrases against the call for papers of the workshops.

Our chief contributions here are the following:

— We propose the exploitation of single-document keyphrase extraction techniques for multi-document keyphrase extraction.

— We evaluate the relative performance of two systems compared against human judgement and a novel automatic procedure as well.

2 Related Work

While single-document keyphrase extraction has been quite well studied, less research has been conducted on multi-document keyphrase extraction. The standard approach is to rank each n-gram in the document set via the information gain metric or χ^2 metric [8]. An extension of an information theory-based metric was introduced in the patent [12], which uses partial mutual information for the determination of keyphrases. Our solution also has an information theoretical basis, but our chief contribution here is that our system exploits deeper positional, linguistic and semantic information concerning the occurrences of the candidate phrases via single-document keyphrase extraction techniques.

Probably the most closely related work to ours is the CorePhrase algorithm [6], which was designed to extract keyphrases for document collections relying on a graph structure, called Document Index Graph. Although the authors of [6] also focused on multi-document keyphrase extraction, their main assumption was that "keyphrases exist in the text and are not automatically generated". Here we focus on the more frequent and realistic case where there are no manually assigned keyphrases available for the documents.

As regards single-document keyphrase extraction, there has been a steady growth of interest, since the pioneering papers of [16] and [14]. Most of the previous papers focused on the domain of scientific papers. A useful benchmark for this particular task is a recently organized SemEval shared task [7] where 19 teams developed keyphrase extractor systems. It is interesting to note that besides the scientific domain, there have been studies on the extraction of keyphrases from different genres of text, e.g. from news articles [15, 4, 3] and product reviews [2].

Keyphrase extraction when performed on (scientific) documents may also be beneficial for research on (scientific) trend detection, as in [5], where the changes in *focus*, *technique* and *domain*-related expressions of scientific publications in the field of computational linguistics were analysed over time. Our study follows this line of research as keyphrases describing a cluster (document set) can provide clues for trend detection.

The growing academic interest in the analysis and processing of scientific literature is reflected by the fact that an entire workshop [1] was devoted to it on the 50^{th} anniversary ACL conference. In that workshop, the authors of [11] introduced the corpus on the previous ACL proceedings, which served as a basis for our experiments.

3 System Description

Our general multi-document keyphrase extraction framework consists of two stages. First, we extract and filter candidate phrases that might represent some subcorpus, then we rank these candidates. We compared two approaches with each other, which essentially differ in the first stage:

— A state-of-the-art-style system which considers all the elements in the union of all the keyphrase candidates of all the documents of the given cluster (referred to as *Baseline*)

— A system which performs single-document keyphrase extraction prior to multi-document keyphrase extraction and which utilizes only the top-ranked single-document keyphrases to determine keyphrases for the given cluster. (This system is abbreviated as *SDK* later on, as it utilizes Single-Document Keyphrases.)

3.1 Candidate Selection and Representation

Candidate selection plays a key role in multi-document keyphrase extraction. Similar to other studies in single-document keyphrase extraction, n-grams consisting of 1 to 4 consecutive tokens were viewed as potential keyphrases if they did not start or end with any element of a predefined set of stopwords. An additional constraint for candidate phrases was that all of their constituting non-stopword tokens had to be identified as *noun*, *adjective* or *verb* by the POS tagger described in [13].

In addition, any n-gram in order to remain on the list of candidate phrases had to fulfil the requirement of having at least one occurrence besides the *References* section of a paper. The reason for this was that phrases that occur just in references are mostly improper keyphrases, such as the phrase *Digital Library*. Improper keyphrases that were easily recognizable even by their surface forms, like those containing non-English characters and those being shorter than 3 characters, were omitted from the candidate list. This kind of reduction step favored the baseline system which, unlike the other approach, did not employ any semantical ranking and pre-selection of the keyphrase candidates and often treated rare but topically unrelated tokens as highly discriminative and thus were worthy of being selected as multi-document keyphrases.

Next, the normalization of the candidates was carried out in a similar way as before; i.e. the canonized representation of a keyphrase candidate lacked any stopwords, and the lower-cased stems of the resulting non-stopword tokens were placed in alphabetical order. This kind of normalization made it possible to treat two n-grams of different surface forms but similar semantics, such as *Innovation diffusion* and *diffusion of innovation*, as equivalent.

3.2 Single-Document Keyphrase Extraction System

We utilized the NUS Keyphrase Corpus [10] and the database of the SemEval-2 shared task on scientific keyphrase extraction [7] as training data for our supervised keyphrase candidate ranker. Our keyphrase ranking solution was based on the posterior probability of a "keyphrase or not" binary MaxEnt model (MALLET [9] implementation). The classifier employs a rich feature set that will be introduced below.

For the feature representation of a candidate keyphrase, first the basic features of tf-idf and relative first occurrence – as described in KEA [16] – were employed. After considering just the first occurrence of a candidate, all of its locations within a document were taken into account by the feature that was assigned the value of the standard deviation of the various document positions of a candidate phrase, yielding high scores for those phrases which were mentioned throughout an entire document.

Owing to the fact that scientific keyphrases tend to have characteristic character suffixes like *-ics*, *-ment* and *-al*, features were generated from the character suffixes of the individual tokens of keyphrase candidates. As knowing the position where a character suffix can be found within an n-gram might also be helpful, we also incorporated into the features whether a certain 2 or 3-gram character suffix was located inside, at the beginning or at end of a phrase candidate. However, the character suffix feature of one token long keyphrase candidates were treated separately. For instance, the features induced by (and thus assigned with a true value) for the candidate phrase *dynamic semantics* are *B-mic*, *B-ic*, *E-ics*, *E-cs*. Named Entity and Part-of-Speech tags of the individual tokens of a candidate phrase were employed in a similar fashion; i.e. including their within-candidate position in the feature space. For instance, for the phrase *dynamic/JJ semantics/NN*, features *B-JJ* and *E-NN* were set to true to indicate that the phrase had commenced with an adjective and ended with a noun, whereas the 1-token-long phrase *semantics/NN* was set to true only for the feature *S-NN*.

Wikipedia (dump file 2011-01-07) was also utilized for feature computations. First, a list of multi word expressions (MWEs) was collected from it. Second, the results of the following tests for a candidate phrase were included in the feature vectors representing them:

— The phrase itself can be found in the list, e.g. *maximal social welfare ratio*

— It was composed of other elements from the list, e.g. *resource allocation problems*, as the phrases *resource allocation* and *allocation problems* were present in the list, but not as a single phrase

— It may be a superstring of an element from the list, e.g. *general analysis remains*, due to the presence of *general analysis* in the MWE list.

Besides the MWEs gathered from Wikipedia, its category hierarchy was utilized as well; i.e. the nominal parts of the anchor texts of the category links of a Wikipedia article that had the same title as a candidate expression were included in the feature set. This way, semantic knowledge was also incorporated in the feature space assigned to a candidate phrase.

3.3 Multi-Document Keyphrase Extraction System

The keyphrases of a cluster are the top-ranked candidates by the information gain metric (the cluster against the rest of the whole corpus). This was calculated for all the candidate phrases of the documents that belonged to the cluster in question. In the case of the baseline system, all the phrases that could function as candidates for the single-document keyphrase extraction system were treated as potential candidates for the entire subcorpus, whereas in the more sophisticated system each document from a subcorpus contributed only with its top-15-ranked keyphrases derived from the single-document keyphrase extractor.

We decided to choose the top-15 keyphrases per documents as the performance of keyphrase extractors tend to fall well below this threshold (i.e. most of the proper keyphrases are included within the top-15 keyphrases of automated systems and simultaneously, false positive predictions are more prevalent above this threshold) . This latter strategy yields a maximum number of $15|D_i|$ potential keyphrase candidates for the i^{th} cluster D_i of size $|D_i|$ in the unlikely case that all the top-ranked keyphrases of the individual documents in D_i were distinct. This theoretical scenario was unlikely as documents within the same cluster share some common topics, which makes it likely that the individual documents of a cluster share at least some keyphrases.

Then, for a subset of the document collection, the top-3 highest ranked candidates based on their information gain – which had at least a high relative frequency within the documents in the particular cluster as the relative frequency of the phrase outside the cluster – were treated as the keyphrases of the given cluster.

4 Experiments and Results

Now we present the dataset that was used in our experiments on multi-document keyphrase extraction and the results achieved with the Baseline and SDK systems.

4.1 Dataset

As we wished to find a way to assign keyphrases to thematically coherent document sets in some corpus, we decided to focus on that part of the ACL Anthology Corpus described in [11] which just contains ACL workshop papers. The reason why we involved workshop papers in our experiments was due to the fact that conference workshops tend to be inherently homogeneous in their topic selection; i.e. they tend to focus on some particular, clearly distinguishable area of the larger scientific community, such as *parsing*, *machine translation* or *sentiment analysis*.

However, there were workshops that we felt important to remove as their areas of interest were too broad to view the papers that were accepted as one coherent set of documents with respect their topics. The elimination of workshops from the database included the kind of proceedings like those of *Empirical Methods of Natural Language Processing* (also known as *EMNLP*), which used to be listed earlier among workshops in the ACL Anthology and which has a topic coverage that is too heterogeneous. The papers suggested for omission were determined by two computational linguistic experts whose inter annotator agreement in terms of accuracy and κ-statistics was 94.6% and 0.667, respectively, which is to be regarded as strong agreement.

4.2 Human Evaluation

Owing to the fact that all the proper keyphrases of a workshop would be difficult (or even impossible) to be listed exhaustively and simultaneously, there existed several ways of expressing semantically equivalent concepts. Hence, we thought that probably the best way of evaluation was to rely on domain experts' knowledge when determining the usefulness of a given set of keyphrases assigned to a workshop.

As a result, 4 researchers from the NLP field were hired to make decisions for each pairs of sets of keyphrases that were assigned to the

Table 1. Statistics of the workshops present in the ACL Anthology Corpus taken from the 6-year timespan that our experiments focused on

Total workshop papers	1946
Total distinct workshops	125
Total workshop papers excluded	411 (21.12%)
Total workshops excluded	15 (8.00%)
Average papers per (non-excluded) workshops	13.95 ± 8.10

workshops being held between years 2000 and 2005 (inclusive) in the database. The manually evaluated subcorpus consisted of 110 workshops, incorporating a total number of 1,535 documents from the entire corpus, as can be seen in Table 1.

Annotators were given the top-3-ranked keyphrases for each workshop taken from both of the system outputs [1] and given the name of the workshop in question (e.g. *ACL-SIGLEX Workshop on Deep Lexical Acquisition*), they had to make one of the following decisions:

— *Positive draw* or D^+ when both sets of keyphrases might be equally helpful in finding a particular workshop as the keyphrases returned are closely related to the topics of that workshop

— *Negative draw* or D^- when both sets of keyphrases are of no use; that is, neither of them would be helpful at all if they were looking for the particular workshop

— *Win* when they are confident that one of the sets of keyphrases would be more helpful if they were to look for the workshop in question.

Note that *Win* decisions were later automatically split into two further subcategories Win_{SDK} and Win_{BL}, depending on whether an annotator viewed the keyphrase output of the single-document keyphrases-based or the baseline system as better, respectively.

In order to create a final assessment of judgements, the individual decisions of the annotators were merged by simply choosing their most frequent decision for each workshop. There were only 9 cases of ties when trying to decide the majority annotation simply by counting, where final decisions were made by revisiting those test cases. This way a final assessment of decisions was determined for each of the 110 human-evaluated workshops based on the independent decisions of 4 human expert annotators. The agreement rates of the four annotators against their combined decisions are listed in Table 2, while Table 3 contains the distribution of the annotation decisions for each annotator and the combined annotation as well. Due to the commonly accepted interpretation of κ-statistics, the annotators' agreement rates is to be regarded as either moderate or substantial.

Table 2. Annotator agreement rates against the final assessment annotation decisions

	Accuracy	κ-statistic
Annotator_1	80 (80.0%)	0.65
Annotator_2	91 (82.7%)	0.69
Annotator_3	75 (68.2%)	0.48
Annotator_4	74 (67.3%)	0.44

It can be seen in Table 3 that taking the majority of the annotators' decisions was useful in the sense that decisions became less ambiguous as D^+ (i.e. tie annotations) almost entirely disappeared. In the same table we also notice that due to the final assessment of annotations, the single-document keyphrase-based multi-document keyphrase outputs were viewed as better compared to the outputs of the baseline system for over 60% of the workshops. The keyphrases assigned for sample workshops by both the baseline and the more sophisticated method are shown in Table 4, which also seems to be consistent with the human evaluation results; that is, the keyphrases assigned by the baseline system are of lower quality compared to the system which assigns keyphrases to sets of documents based on the individual keyphrases of the documents that comprise the document set.

[1] Annotators were not told which set of keyphrases was determined by which approach so as to reduce the possibility of bias in the annotation procedure. Also the order of the two system outputs was randomized from workshop to workshop.

Table 3. The class distribution of the annotation types of the individual annotators and that of the merged final assessments

	D^+	D^-	Win_{SDK}	Win_{BL}
Annotator$_1$	8 (7.3%)	12 (10.9%)	63 (57.3%)	27 (24.5%)
Annotator$_2$	4 (3.6%)	12 (10.9%)	62 (56.4%)	32 (29.1%)
Annotator$_3$	19 (17.3%)	9 (8.2%)	55 (50.0%)	27 (24.5%)
Annotator$_4$	17 (15.5%)	16 (14.5%)	54 (49.1%)	23 (20.1%)
Final assessment	1 (0.9%)	10 (9.1%)	**69 (62.7%)**	30 (27.3%)
Automatic evaluation	0 (0.0%)	18 (16.4%)	49 (44.5%)	43 (39.1%)

Table 4. Simple outputs of the two approaches for various workshops taken from the ACL Anthology Corpus that have the baseline keyphrases and the single document-based keyphrases on the left and right hand sides, respectively

Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization	
fluency	*machine translation evaluation*
automatic scores	*automatic evaluation*
rouge	*MT evaluation*
Multilingual Question Answering	
correct answer	*QA system*
answer type	*question answering*
monolingual systems	*answering system*
Information Retrieval with Asian Languages	
term frequency	*information retrieval*
retrieval system	*representative keywords*
document frequency	*semantic indexing*
Web as Corpus	
web corpus	*Web as corpus*
wacky project	*search engine*
wacky	*corpus data*

Table 5. The growth in the number of distinct multi-document keyphrase candidate word forms as a function of processed documents.

Documents processed	500	1000	1535	2500	5000
Baseline	88,732	150,254	206,883	292,063	485,982
SDK	4,270	7,582	10,665	15,817	27,739

4.3 Automated Evaluation

Besides human evaluations, another automated evaluation was carried out. These experiments were conducted on the same workshop data as those for the human evaluations; i.e. the ones that were held between the years 2000 and 2005 (inclusive) and were not judged to be too general in their topic. In order to measure the quality of the workshop-level keyphrases, the original call for papers (CFPs) of the workshops were crawled from the Web, the contents of which served as the basis of comparison for the extracted workshop-level keyphrases.

We should mention here that other methods, besides relying on the original CFPs for the workshops, were experimented with, like assigning Wikipedia articles (e.g. *Natural Language Generation*) to workshops according to their topics and examining the overlap between the extracted keyphrases of a workshop and the contents of the Wikipedia article from the same workshop it was assigned to. However, we found that several areas of NLP lacked any truly relevant Wikipedia article that could be assigned to it and even those topics that had a Wikipedia article, the degree of elaborateness was markedly different across the various communities of NLP.

Using the basic information retrieval techniques described in [8], the quality of each system was measured in the following way. Two vectors were created for the two approaches, both incorporating dimensions for the 1,-and 2-grams of those phrases that could be regarded as keyphrase candidates (as described in Section 3.1) of the call for papers. For the automatic decision of which systems' output should be regarded as better for a particular workshop, two meta-document vectors were created for the two systems, having non-zero entries just for the top-3 keyphrases. These meta-documents functioned as query vectors and the one which had the greater cosine similarity to the CFP-based prototype vector of the given workshop was selected.

In order to prioritize via term importance within the documents, a tf-weighting of the phrases in the vector space was used. For the workshop-level meta-vectors, the *tf* term was calculated as the weighted relative frequency of the candidates across all the documents belonging to the workshop. Expressed in formal terms, the baseline method was preferred for a workshop i if

$$\frac{x^{\intercal}_{CFP,i} x_{baseline,i}}{\| x_{CFP,i} \| \| x_{baseline,i} \|} > \frac{x^{\intercal}_{CFP,i} x_{SDK,i}}{\| x_{CFP,i} \| \| x_{SDK,i} \|}. \quad (1)$$

In this kind of evaluation, D^- decisions were equivalent to the situation where neither of the top-3 ranked keyphrases intersected the CFP-based prototype vector, thus resulting in a 0 similarity. In the last row of Table 3, we see that this latter kind of evaluation was obtained more frequently by the automatic method than by humans, but we should add that a keyphrase that is not present in the CFP of a workshop is not necessarily useless for a given workshop. Equal but non-zero similarities would have yielded D^+ annotations for workshops, but this situation never occurred. In the remaining cases, the Win_{SDK} decision was obtained during the automated evaluation phase.

4.4 Efficiency

The method that we proposed – i.e to rely just on the best-ranked document-level keyphrases and not on all the keyphrase candidates of the individual documents when performing keyphrase extraction for multiple documents – has various advantages. Not only is the quality of the keyphrases superior compared to the baseline approach based on the human and automatic evaluations, but from Table 5 it is also clear that the size of the vocabulary from which keyphrases of document subsets are finally selected can also be reduced by several orders of magnitude even for a corpus of a few thousand documents. Smaller vocabulary naturally makes multi-document keyphrase extraction less resource-intensive and faster without any loss in the quality of keyphrases produced.

5 Conclusions and Future Work

Our results that sought to assign workshop-level keyphrases for the ACL Anthology Corpus suggest that single-document keyphrase extraction can enhance the effectiveness in multiple-document keyphrase extraction. Both human and automatic evaluations on a 6-year time slot of the corpus show a superior quality over the baseline system. We think that similar efforts carried out on scientific archives can support scientific communities.

In the future, we would like to examine the possible use of document subset-level keyphrases in the detection of similar topics within text corpora and trend analysis. The use of document- and subcorpus-level keyphrases should be beneficial for document set visualization, which we would like to verify as well in the future.

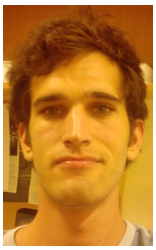
Acknowledgments

This work was in part supported by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

References

1. **Banchs, R. E.**, editor (**2012**). *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Association for Computational Linguistics, Jeju Island, Korea.
2. **Berend, G.** (**2011**). Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 1162–1170.

3. **Ding, Z., Zhang, Q., & Huang, X.** (**2011**). Keyphrase extraction from online news using binary integer programming. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 165–173.
4. **Farkas, R., Berend, G., Hegedűs, I., Kárpáti, A., & Krich, B.** (**2010**). Automatic free-text-tagging of online news archives. In *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*. IOS Press, Amsterdam, The Netherlands, The Netherlands. ISBN 978-1-60750-605-8, 529–534.
5. **Gupta, S. & Manning, C.** (**2011**). Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 1–9.
6. **Hammouda, K. M., Matute, D. N., & Kamel, M. S.** (**2005**). Corephrase: keyphrase extraction for document clustering. In *Proceedings of MLDM*. 265–274.
7. **Kim, S. N., Medelyan, O., Kan, M.-Y., & Baldwin, T.** (**2010**). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10. ACL, Morristown, NJ, USA, 21–26.
8. **Manning, C. D., Raghavan, P., & Schtze, H.** (**2008**). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. ISBN 0521865719, 9780521865715.
9. **McCallum, A. K.** (**2002**). Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.
10. **Nguyen, T. D. & Kan, M.-Y.** (**2007**). Keyphrase extraction in scientific publications. In *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers*, ICADL'07. Springer-Verlag, Berlin, Heidelberg. ISBN 3-540-77093-3, 978-3-540-77093-0, 317–326.
11. **Schäfer, U., Read, J., & Oepen, S.** (**2012**). Towards an acl anthology corpus with logical document structure. an overview of the acl 2012 contributed task. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Association for Computational Linguistics, Jeju Island, Korea, 88–97.
12. **Surendran, A. C.** (**2010**). Multi-document keyphrase extraction using partial mutual information. Patent. US 7711737.
13. **Toutanova, K. & Manning, C. D.** (**2000**). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, EMNLP '00. ACL, Stroudsburg, PA, USA, 63–70. doi:http://dx.doi.org/10.3115/1117794.1117802.
14. **Turney, P.** (**2000**). Learning algorithms for keyphrase extraction. *Information Retrieval*, 2, 303–336.
15. **Wan, X. & Xiao, J.** (**2008**). Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, AAAI'08. AAAI Press. ISBN 978-1-57735-368-3, 855–860.
16. **Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Craig** (**1999**). Kea: Practical automatic keyphrase extraction. In *ACM DL*. 254–255.



Gábor Berend is a researcher at the Natural Language Processing Group of the University of Szeged. The title of his thesis (to be submitted) is "Machine Learning-based Extraction of Keyphrases and its Applications in Various Domains". He has been working with natural language processing problems at the Artificial Intelligence Research Group of the University of Szeged since 2009. His main interests of research are information retrieval and extraction.

Richárd Farkas is a senior researcher at the Natural Language Processing Group of the University of Szeged. He received his PhD from the University of Szeged in 2010 (title of the thesis: "Machine Learning techniques for applied Information Extraction"). He has been working as a postdoctoral researcher in the group of Hinrich Schütze at the Institute for Natural Language Processing, University of Stuttgart between 2011 and 2012. His main interests of research are parsing morphologically rich languages and non-canonical texts; the syntactic parse-relation extraction interface; real world applications of statistical natural language processing.

Article received on 07/12/2012; accepted on 13/01/2013.