



UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
<http://www.disi.unitn.it>

LARGE DATASET FOR KEYPHRASE EXTRACTION

Mikalai Krapivin, Aliaksandr Autaeu, Maurizio Marchese

May 2008

Technical Report # DISI-09-055

Large Dataset for Keyphrase Extraction

Mikalai Krapivin
Department of Information
Engineering and Computer Science
University of Trento, Italy
krapivin@disi.unitn.it

Aliaksandr Autaeu
Department of Information
Engineering and Computer Science
University of Trento, Italy
autaeu@disi.unitn.it

Maurizio Marchese
Department of Information
Engineering and Computer Science
University of Trento, Italy
marchese@disi.unitn.it

ABSTRACT

We propose a large dataset for machine learning-based automatic keyphrase extraction. The dataset has a high quality and consist of 2,000 of scientific papers from computer science domain published by ACM. Each paper has its keyphrases assigned by the authors and verified by the reviewers. Different parts of papers, such as title and abstract, are separated, enabling extraction based on a part of an article's text. The content of each paper is converted from PDF to plain text. The pieces of formulae, tables, figures and LaTeX mark up were removed automatically. For removal we have used Maximum Entropy Model-based machine learning and achieved 97.04% precision. Preliminary investigation with help of the state of the art keyphrase extraction system KEA shows keyphrases recognition accuracy improvement for refined texts.

1. INTRODUCTION

Scientific Digital Libraries present a lot of challenges in different unsupervised or semi-supervised information extraction problems. Modern Digital Libraries like CiteSeerX¹ or Google Scholar² contain millions of documents. Typically, the crawler downloads a document, converts it to a plain text format and then extracts all necessary information. Meta-information like author name, affiliation, title is a kind of explicit information that with a great probability should be inside a text. There are corpora for such information extraction tasks [11], and there are papers describing such extraction methodologies [11][2]. A greater challenge is to extract information which is implicit, like concepts or keyphrases [7][16][17]. There are papers describing different methodologies for keyphrase extraction, for instance, pioneering work of Tourney [15] or more recent papers [3]. The state of the art contains complaints about absence of standard benchmarking sets for keyphrase extraction validation and methodology proof [9]. In this paper we address this problem and present a large good quality dataset for keyphrase extraction. We hope it will establish a ground for fair evaluation and comparison of different keyphrase extraction systems.

2. Problems with existing datasets

Let us briefly mention some previous works about keyphrase extraction from the point of view of benchmarking set usage. Chronologically the pioneer in successful keyphrase extraction was Peter Tourney [15]. He proposed very detailed investigation of decision trees based algorithms and several links to freely available datasets. For instance, NEXOR³, FIPS⁴ and

others (see [15] for more complete sets description). However, that was more than decade ago and *all* those links no longer available and we have failed to find any of the proposed datasets in internet. Later work which is one of the most valuable in the domain is KEA⁵ [17]. Its algorithm is based on Naive Bayes classifier. KEA is a free software and can be downloaded through KEA website, but there are no standard datasets in the download package. KEA inventors mention that they obtained Tourney dataset directly from the author. Nguen et al [9] directly pointed out to the impossibility to find any proper datasets and used their own dataset constructed from 250 crawled documents.

Creation of benchmarking sets is not a new field. There are some datasets well-known in Information Retrieval. For example is Reuters Dataset, prepared by David Lewis⁶. This dataset carries thousands of short news texts with labels, and helps to evaluate classification algorithms. Another example is a large dataset called TREC⁷. TREC collection is dedicated to web mining, indexing and query answering. It fits well to semantic search community tasks and has been used in different semantic-based and NLP evaluations.

3. Dataset types

We emphasize that most of the datasets proposed in state of the art belong to the area of scientific papers. For instance, Tourney [15] used **75** scientific papers from different domains: Neuro Science, Behavioral and Brain Science and Chemistry on one hand. He also used **311** email messages and up to **140** of web pages from different domains. Dataset of a similar size was mentioned in [1]. In the more recent work Tourney proposed **500** of scientific papers from Physics domain taken from

arXiv.org e-Print archive⁸. Papers were taken in PostScript (PS) format and author did not mention neither how they converted them to text nor what is the conversion quality. In the [1] authors proposed a dataset consisting of **160** scientific papers without mentioning particular domains. Annette Hulth [4] took **198** pieces of short Swedish texts related to social activities. In previous work she proposed commercial dataset from Inspec⁹ [5]. We have recently proposed a novel method combining state of the art

⁴ <http://www.itl.nist.gov/div897/pubs/>

⁵ <http://www.nzdl.org/Kea/>

⁶

<http://www.daviddlewis.com/resources/testcollections/reuters21578/>

⁷ <http://trec.nist.gov/>

⁸ <http://arxiv.org/>

⁹ <http://www.theiet.org/publishing/inspec/>

¹ <http://citeseerx.ist.psu.edu/>

² <http://scholar.google.com>

³ <http://www.nexor.com/public/aliweb/search/doc/form.html>

Support Vector Machines learning in combination with Stanford NLP Parser [7] upon 400 of scientific papers.

3.1 Dimensionality

Dimensionality is one of the biggest problems of any keyphrase extraction research papers. To the best of our knowledge no one used dataset containing more than 500 scientific papers. We think this is caused by the difficulties in dataset construction and further results evaluation. Most of the trials were done *manually*, which is extremely time consuming. However, increasing the dataset size may lead to significant improvements in precision and recall of tasks based on supervised machine learning methods (See Section 5).

3.2 Reproducibility

Reproducibility is the main problem of most of the datasets considered in the state of the art. It is really hard to compare new algorithms and methodologies with all previous works since the results may vary from dataset to dataset drastically. Even taking papers from the same storage and nearly same domain may change all the results depending on machine learning method.

4. Dataset description

The dataset we present contains all papers from Computer Science domain, which were published by ACM in the period from 2003 to 2005. All these papers are written in English and stored in UTF-8 text encoding. Each text has clearly indicated:

- Title
- Abstract
- Body
- References (recognized by our method [6])
- References crawled from ACM portal
- References to citing papers (also taken from ACM)

The separation of the parts enables to use them as an additional training material for training text part recognition. Moreover, they can be used to restrict search for a keyphrase to a part of the text. For example, search can be restricted to abstract and references only, as it was done in [7][16]. This is convenient for not very scalable methods like SVM [16]. Each file holds full text of a paper and has the name like "[id].txt" where "[id]" is a valid ACM¹⁰ document id, for instance "1005858.txt" corresponds to a real paper with id equal to "1005858". One may find this paper at <http://portal.acm.org> and make sure it is a paper "A framework for architecting peer-to-peer receiver-driven overlays" with attached keyphrases "congestion control, peer-to-peer streaming". Keyphrases for particular file are located in file "[id].key". This format is used in KEA [17]. Dataset contains precisely 2304 papers freely available in internet¹¹. It is not separated to a training set and a test set, so we presume applying of cross-validating procedure (see for example [10]).

The papers full texts were downloaded from CiteSeerX Autonomous Digital Library.

¹⁰ <http://www.acm.org/>

¹¹ <http://disi.unitn.it/~krapivin/>

5. Dataset preparation

We took the papers in PDF format from CiteSeerX, skipping all corrupted or "unconvertible" PDFs (such as PDF stored as image). Metainformation like titles, references and abstracts was taken from ACM portal. We have mapped ACM metainformation to CiteSeer texts on the bases of crawled id mappings and information kindly shared with us by Professor Lee Giles, Pennsylvania State University (CiteSeer, CiteSeerX). Then we converted PDF to plain text using a commercial system, than information was processed step by step as described in [6]. Doing this we have found some "garbage", or lexically meaningless pieces of information, which trapped into texts as a result of double conversion: from LaTeX to PDF and then from PDF to text. While using Natural Language Processing tools may improve keyphrase recognition rate [7][4] this "garbage" decreases the precision of Natural Language Processing parsers. We have used Maximum Entropy Model-based training to eliminate the "garbage".

Garbage cleaning

PostScript and PDF formats are current standard of presenting scientific papers. While they have many advantages of allowing rich formatting, complex formulas and figures to be used, for many tasks requiring natural language processing this presents an additional challenge of extracting plain text out of a PDF document. Many tools address the issue of PDF to plain text conversion. However, the resulting plain text document often contains remains of LaTeX markup, various extra punctuation symbols, clusters of brackets. For example, example below shows the example of a "garbage" remaining in plain text after the conversion from PDF.

a linear system $Ax = b$, in which
satisfy $k = (M \setminus \Gamma_{m1} N)$, so the
iteration

These markup and punctuation pieces restrain modern NLP tools from achieving maximum performance and even cause failures in less robust tools. Therefore, it is desirable to clean up this "garbage" from the text. Example below (in big font) shows how cleaned text looks like. Cleaned in this way text eliminates failures in NLP tools and allows them to achieve better results.

a linear system b , in which
satisfy, so the iteration

Due to the large size of the dataset, manual cleaning will take a lot of time and is unfeasible. Our approach is to use supervised machine learning. The task of identifying the garbage in a text could be seen as deciding for each token its category, which could be either "text" or "garbage". We annotate a small sample of the dataset, consisting of 6 documents containing together about 53,000 tokens. To each token we attach a tag identifying whether it is a "text" token or a "garbage". The task of classifying text tokens into different categories is well-known in NLP as part-of-speech tagging. We train and evaluate two state-of-the-art part-of-speech taggers, Stanford POS tagger [14] and OpenNLP tools [8] POS tagger on our annotated dataset. Both of them are based on Maximum Entropy Models [12]. We tried several combinations of options available in taggers, however the best performance was achieved using default settings.

For tagging we use approach described in [13]. We use our own very small tag set of 2 tags, namely T for text and G for garbage. We extract and use the following features to make tagging decisions:

- up to 4 prefixes made of first 4 characters
- up to 4 suffixes made of last 4 characters
- presence of punctuation characters inside a token
- presence of initial capital letter in a token
- presence of digits inside a token
- 2 previous words and their tags
- 2 successive words and their tags

We evaluate both taggers using 10-fold cross-validation on our annotated sample. Table 1 summarizes taggers performance.

	Precision per token, %	Garbage, %
Stanford	95.55	83.19
OpenNLP	97.04	87.21

Table 1. POS Taggers performance

For the better performing OpenNLP POS tagger Figure 1 shows precision improvement during incremental training

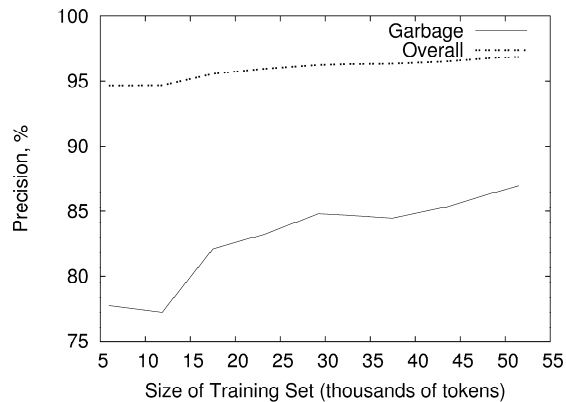


Figure 1. Incremental training. Dashed line shows precision over all tags. Solid line shows percent of removed garbage.

Note the stabilization of the overall precision curve of the POS tagger around 97%. While the overall precision stabilizes, we note that solid line showing precision per tag G, which indicates "garbage" to remove, is not stable yet. This precision might be improved further by increasing the size of our manually annotated training set.

5.1 Preliminary evaluation

To evaluate how much text refinement may affect precision of keyphrase extraction, we have performed few preliminary experiments. We took state of the art system KEA [17] and performed the extraction upon 120 arbitrary selected documents.

We split the dataset into 2 parts, a training set and a testing set. After that we trained KEA Naive Bayes based model with 2-fold cross validation. Table 2 summarizes the results.

Dataset	F-Measure, %
Refined texts	19.08

Table 2. Preliminary comparison of keyphrase extraction for refined and unrefined texts.

We think the improvement is small because KEA does not use a lot of syntactic information. However, KEA sometimes recognizes special symbols or a piece of a formula as a keyphrase, and therefore it performs better on refined texts, which contain much less noise of this kind.

Using refined texts improves even performance of simple keyphrase extraction methods which do not use syntactic information. We expect greater performance boost for methods using advanced NLP techniques.

Conclusion

We have prepared and presented a large dataset for keyphrase extraction. The novelties of the dataset are:

- It is at least 10 times bigger than all previously used datasets.
- It is a set of full texts of scientific papers, which is typical for the keyphrase extraction domain.
- It has author assigned and editor corrected keyphrases.
- It is verifiable, because all presented information may be found through CiteseerX and ACM portal.
- It is public and available for use by researchers.
- It is refined for better NLP processing to get more syntactical and semantic knowledge.

The proposed dataset may also be used for classification tasks, because all presented documents have classification labels which may be found on ACM portal. Another possible use of the dataset is the text parts detection tasks.

6. ACKNOWLEDGMENTS

Authors acknowledged Prof. Lee Giles for the presented papers collection.

7. REFERENCES

- [1] C. Ercan and I. Cicekli. Using lexical chains for keyword extraction. 43:1705-1714, 2007.
- [2] H. Hui, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, , and E. A. Fox. Automatic document metadata extraction using support vector machines. In Proceedings of the 3rd ACM/IEEE-CS JCDL), pages 37-48, 2003.
- [3] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing), volume 10, pages 216-223, 2003.
- [4] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing), volume 10, pages 216-223, 2003.
- [5] A. Hulth, J. Karlgren, A. Jonsson, H. Bostrom, and L. Asker. Automatic keyword extraction using domain knowledge. 2004.

- [6] A. Ivanyukovich and M. Marchese. Unsupervised free-text processing and structuring in digital archives. In 1st International Conference on Multidisciplinary Information Sciences and Technologies, 2006.
- [7] M. Krapivin, M. Marchese, A. Yadrantsau, and Y. Liang. Unsupervised key-phrases extraction from scientific papers using domain and linguistic knowledge. In Digital Information Management. ICDIM 2008), pages 105-112, London, GB, Nov 2008.
- [8] T. Morton. Using Semantic Relations to Improve Information Retrieval. PhD thesis.
- [9] T. D. Nguyen and M. Yen Kan. Keyphrase extraction in scientific publications.
- [10] E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. In Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR 97), page 130, Washington, DC, USA, 1997. IEEE Computer Society.
- [11] F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. In Proceedings of Human Language Technology Conference, 2004.
- [12] A. Ratnaparkhi. A maximum entropy part of speech tagger. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. University of Pennsylvania, 1996.
- [13] A. Ratnaparkhi. A simple introduction to maximum entropy models for natural language processing. Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania, Aug. 1997.
- [14] K. Toutanova and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In EMNLP/VLC 2000, pages 63-70, 2000.
- [15] P. Turney. Learning to Extract Keyphrases from Text. Technical report, NRC/ERB-1057, Feb. 1999.
- [16] J. Wang and H. Peng. Keyphrases extraction from web document by the least squares support vector machine. In IEE/WIC/ACM International Conference on Web Intelligence), 2005.
- [17] I. H. Witten, G. W. Payne, E. Frank, C. Gutwin, and C. G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In Proceedings of DL T99, pages 254-256, 1999. <http://www.nzdl.org/Kea/>.