Marek KOZŁOWSKI
NATIONAL INFORMATION PROCESSING INSTITUTE
Al. Niepodległości 188 B, 00-608 Warszawa

PKE: a novel Polish keywords extraction method

M.Sc. Marek KOZŁOWSKI

M.Sc. in computer science (2008), M.Sc. in economy (2011). I am scientific researcher at National Information Processing Institute (OPI) and assistant at Warsaw University of Technology. In OPI I am responsible for leading a team of engineers working on recommendation and analytic systems for education and research market. I am leading the projects of data acquisition and analysis for e.g: Millward Brown, NCBR I am developing computational methods (namely text and web mining methods).

e-mail: mkozlowski@opi.org.pl

Abstract

In the paper a novel summarization approach, called the Polish Keywords Extractor (PKE), is presented. It is the single document oriented method that is capable of extracting keywords from Polish documents. PKE is a knowledge-poor method (not using any external knowledge resources as Wikipedia) inspired by RAKE and KEA. In comparison with the previous methods PKE uses Polish lemmatizer, Part-of-Speech filters, and various evaluation approaches (statistical measures, classifiers). This algorithm was tested on a set of abstracts of Polish academic papers. The experiments have shown that PKE achieves better quality measures (precision, recall, F-measure) than RAKE and KEA.

Keywords: information retrieval, keyword extraction, summarization.

PKE: nowatorska metoda ekstrakcji słów kluczowych dla języka polskiego

Streszczenie

Automatyczne streszczanie tekstów dotyczy redukcji całych dokumentów lub korpusów dokumentów do postaci reprezentatywnego zbioru słów, lub akapitu. Jedną z popularniejszych metod streszczania jest ekstrakcja słów kluczowych, której celem jest identyfikacja pojedynczych słów lub fraz etykietujących zadany dokument. Metody ekstrakcji słów kluczowych mogą być podzielone na zorientowane na pojedyncze dokumentu lub na korpusy. Dodatkowo metody ekstrakcji mogą być klasyfikowane według stosowanych podejść: lingwistyczne podejście, statystyczne lub oparte na uczeniu maszynowym. W tym artykule jest zaprezentowane nowe podejście do ekstrakcji słów kluczowych, nazwane PKE, które jest zorientowane na pojedyncze polsko językowe dokumenty. PKE jest metodą nie wykorzystującą zewnętrznych zasobów wiedzy jak np. Wikipedia. Metoda została zainspirowana metodami KEA [7] i RAKE [8]. RAKE jest algorytmem bez nadzoru, niezależnym od dziedziny i języka, który pozyskuje słowa kluczowe z pojedynczych dokumentów. KEA natomiast jest metodą z nadzorem, która wykorzystuje modele bayesowskie w celu obliczenia prawdopodobieństwa bycia słowem kluczowym. W porównaniu do powyższych metod, PKE używa Polskiego lematyzatora, filtrów części mowy, oraz różnorodnych metod ewaluacji (statystycznych miar, klasyfikatorów). Proponowany algorytm został przetestowany na zbiorze polskich abstraktów artykułów. Automatycznie proponowane słowa kluczowe zostały zweryfikowane względem słów wybranych przez autorów prac. Eksperymenty (tabela 1 i 2) pokazały, że PKE osiąga lepsze miary jakości (precyzja, kompletność, F1) niż RAKE i KEA.

Słowa kluczowe: pozyskiwanie informacji, ekstrakcja słów kluczowych, automatyczne streszczanie.

1. Introduction

Automatic summarization consists in reducing a textual document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text. There are two methods for automatic text summarization - extractive and abstractive. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original. The most of research is focused on extractive methods, and this is what we will cover.

In general, the process of automatic text summarization is divided into three stages [1]: analysis of the given text, summarization of the text, and presentation of the summary in a suitable output form. Two particular types of extractive summarization are addressed in the literature: keyword extraction and sentence extraction. The first method goal is to select individual words or phrases to tag a document. A keyword is either a single word (unigram) or a compound word (n-gram), representing an important concept. The second method goal is to select whole sentences to create a short paragraph summary. In our survey we have focused on keyword extraction from single documents.

Keywords of a document are usually several words or phrases that are closely related to the content of the document. Keywords provide rich semantic information for many text mining applications, e.g: document classification, clustering and topic search. Many academic conferences require that each paper will include the paper title, its abstract and a set of keywords. The keywords provide general information about the contents of the document and can be seen as an additional kind of a document abstraction. In libraries professional indexers select keywords from a controlled vocabulary according to the defined cataloguing rules. Digital libraries, or any repositories of data (flickr, blog articles etc.) also use keywords (or here called *content tags* or *content labels*) to organize and provide a thematic access to their data. The keyword concept is in the Internet used by various search-engines. Due to the simplicity of keywords, search engines are able to handle huge volumes of free text documents.

2. Related work

Methods of keyword extraction can be divided into two groups – single document oriented ones and corpus oriented ones. First approach is focused on discovery meaningful words inside the document in order to index it. The second approach assumes that there is a representative corpus (set of documents) which may be mined in order to get collection of significant concepts and their potential relations. Extracting terms from corpus is very often used in the problem called ontology learning from text. Our survey is constrained to extraction keywords from sa ingle document (precisely abstract).

Keyword extraction methods may be categorized by the type of technique used to identify important words. Major types are linguistic approach and statistical approach.

Linguistic methods are general based on NP (noun phrase) chunks or apriori defined PoS (Part of Speech) patterns. Justeson and Katz [2] discovered relevant terms using the PoS pattern:

$$((Adj|Noun)^{+}|((Adj|Noun)^{*}(NounPrep)^{?})(Adj|Noun)^{*})Noun \quad (1)$$

where:

- *Adj* – adjective;
- *Noun* – noun;
- *NounPrep* – noun preposition.

Daile [3] used also PoS patterns to identify significant terms. Patterns were simply bigrams: noun-noun and adjective-noun.

A statistical approach is based on counting the term frequency. Following this methodology the most important terms are those which are often represented in a text. The first works identified

terms by counting the frequency of single words in the context of the selected corpus. Jones [4] describes a survey on discovering statistical discriminating words in the set of documents. Andrade and Valencia [5] compare the distribution of a candidate word in a document with the distribution of this word in a reference corpus. A disadvantage of the statistical approach is avoiding infrequent words, which could be important ones. Especially in short texts (e.g. abstracts) there is very unlikely to find an important term (e.g. natural language processing) that may be not multiple repeated. Therefore the currently popular approach merges the above mentioned statistical approach with linguistic elements and heuristics.

Anette Hulth [6] presents the extraction method based on a supervised machine learning approach. Treating automatic keyword extraction as a supervised machine learning task means that a classifier is trained by documents with known keywords. Each new candidate term is classified either as a keyword or a non-keyword. Each candidate must be a NP-chunk (noun phrase). The classifier uses four features: term frequency, inverse document frequency, relative position of the document preceding the first occurrence, and PoS tag or its combination. A machine learning algorithm used in classification is rule induction with recursive-partitioning strategy.

The classifier is also exploited in the KEA method. KEA (Keyprashes extraction algorithm) [7] is an algorithm for extracting keyphrases from text documents. It can be either used for free indexing or for indexing with a controlled vocabulary. Before being able to extract keyphrases from new documents, Kea first needs to create a model that learns the extraction strategy from manually indexed documents. Naive Bayes classifier is trained using set of manually labeled documents. Kea extracts n-grams of a predefined length (e.g. 1 to 3 words) that do not start or end with a stopword. In controlled indexing, it only collects those n-grams that match thesaurus terms. For each candidate phrase Kea computes 4 feature values: TF-IDF, first occurrence (terms that tend to appear at the start or at the end of a document are more likely to be keyphrases), length (number of component words), node degree (only in the case of thesaurus). When extracting keyphrases from new documents, Kea takes the Naive Bayes model and feature values for each candidate phrase and computes its probability of being a keyphrase. Phrases with the highest probabilities build the final set of keyphrases.

Opposite to the above mentioned supervised methods RAKE [8] is an unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents. In developing RAKE, the motivation has been to create a keyword extraction method that is extremely efficient, operates on individual documents, is easily applied to new domains, and operates well on multiple types of documents. RAKE is based on the observation that keywords frequently contain multiple words but rarely contain standard punctuation or stop words. RAKE begins keyword extraction on a document by parsing its text into a set of candidate keywords. First, the document text is split into sequences of contiguous words at phrase delimiters and stop word positions. Words within a sequence are assigned the same position in the text and together are considered as candidate keyword. After each candidate keyword is identified and the graph of word co-occurrences is complete, a score is calculated for each candidate keyword and defined as the sum of its member's word scores. There are several metrics for calculating word scores, based on the degree and frequency of word vertices in the graph: (1) word frequency (*freq(w)*), (2) word degree (*deg(w)*), and (3) ratio of degree to frequency (*deg(w)/freq(w)*). Words that predominantly occur in longer candidate keywords are favored by *deg(w)/freq(w)* . The ratio of degree to frequency is the final factor used by RAKE to decide about term importance. After candidate keywords are scored, the top *T* scoring candidates are selected as keywords for the document.

3. The PKE method

PKE (**P**olish **K**eywords **E**xtractor) is a self-made keywords extraction method inspired by RAKE and KEA. In comparison with the original methods it has Polish lemmatizer, Part-of-Speech filters, and different evaluation solutions (function (*freq(w)^2*), or Naive Bayes classifier). The proposed algorithm has two candidate selection methods, and two candidate evaluation methods, which provides us with four possible configurations. All of them are tested in details in the below paragraphs.

PKE starts with splitting a text into sentences. Each sentence is represented as a sequence of words. Next, the words are normalized (lemmatized) and tagged with Part-of-Speech properties. Keyword candidates selection is performed in order to find a finite number of potential significant words. Finally the candidates are evaluated by models (classifier or statistic function), and the limited number of them with the highest scores are retrieved.

Candidate selection approaches

Pattern-Recursive selector is based on four ordered PoS patterns. First, from a sentence there are removed stopwords, words containing none-alphanumeric items, words, which are not nouns, adjectives, or unknown PoS. Next, there is invoked a recursive finder, which searches for candidates using four priority patterns (noun-adjective, noun-unknown, noun-noun, noun). When it identifies the phrase matching one of these patterns then the sentence is splitted and subsentences are recursively searched in the same manner. In this way potential candidates are identified. Additionally, for each candidate there are generated subcombinations (if its size is more than one word), which are checked by the above mentioned PoS patterns. Subcombinations extend list of potential keywords.

PoS selector is based on the pattern described by the complex regular expression. The text is splitted into sequences of contiguous words at phrase delimiters and stop word positions. Next for each candidate there are built its subcombinations, which extend the number of potential keywords. All of the potential keywords are filtered out by the PoS regular expression: *(noun)(noun|adjective|unknown)**.

Candidate evaluation approaches

Evaluation may be performed in two modes – supervised (classifier) and unsupervised (statistic function) one.

The supervised approach is based on classifiers. Each candidate is categorized as a proper or unproper keyword. The candidates are described by vectors of four features (all features are normalized to one):

- tf-idf – term frequency is counted for a candidate in the processed abstract, inverse document frequency is counted within training set;
- first occurrence in an abstract – a position of the first candidate's occurrence in a processed abstract;
- first occurrence in a sentence – a position of the first candidate's occurrence within a sentence;
- size – number of member's words contained in a keyword.

In the training phase the classifier is learned to identify proper keywords. During candidate validation there are measured probabilities of being a proper keyphrase and being unproper one. The highest of them implicates the category. Probabilities of being a proper keyword are used to sort candidates and retrieve only the limited number of them.

The unsupervised candidates evaluation approach is based on scoring function *freq(w)^2*. Compound keyword scores are the sum of its member's scores. This approach influences the compound keyword scores, making it higher than its members. This formula reaches better results than the Rake's one –

deg(w)/freq(w). The scored candidates are sorted and top *T* are selected as keywords for the selected document.

4. Experiments

Keywords, which we define as a sequence of one or more words, provide a compact representation of the document contents. In this study, we evaluate a few methods that are capable of extracting keywords from single documents. We have tested these methods on the set of abstracts of academic papers. All abstracts contain keywords composed by the authors. We compare the automatic proposed keywords to the authors' keywords.

The abstracts were downloaded from different web sources (e.g. pubmed, yadda). All the abstracts have assigned at least three keywords. All manually assigned keywords tend to appear in a text of the abstract (about 83% of keywords are directly placed within text of abstract). The abstracts were divided randomly into a training set (about 9000 documents) and validation set (about 3000 documents). Tests were taken on a validation set.

We choose abstracts instead of full text articles because of two reasons. First of all it is much more easier to automatic download abstracts from digital libraries. Most of digital libraries demand authorization to download a whole article, and there is also legacy issues concerning saving articles on the private hard disks. The second reason is much more pragmatic, there is lots of academic papers about article summarization, but only few of them involve abstracts. You cannot use the same approach for a long text as for a short one, which causes this topic interesting. The vast majority of texts in the web is shorter than longer. We could assume that our results and methods, built in order to retrieve keywords from abstracts, may be very easily imported into web resources.

In the next subsections we analyze RAKE and KEA results on Polish abstracts, next PKE results, and finally we compare them.

4.1. RAKE and KEA Evaluation

KEA and RAKE were tested on the set of Polish abstracts.

In Table 1 a few variants of RAKE and KEA methods are tested. In the column *methods* each algorithm name ends with number – 5, 10 or 15. It means the upper limitation of retrieved keywords. RAKE does not have any phase of learning. KEA has got the classifier learned on our training set of abstracts. Benchmark of methods is based on three measures: precision, recall, and F-measure.

Tab. 1. Results for RAKE and KEA on Polish abstracts (tbe best results in bold)
Tab. 1. Wyniki RAKE i KEA na zbiorze Polskich abstraktów (najlepsze wyniki pogrubione)

methods	Abstracts no	mean no. of extracted keywords	precision	recall	F-measure
RAKE 5	3117	4,991979	1,343188	1,408166	1,37491
KEA 5	3117	5	**18,408726**	**19,330279**	**18,858251**
RAKE 10	3117	9,755213	3,357779	6,879127	4,512807
KEA 10	3117	10	**13,227462**	**27,779275**	**17,921412**
RAKE 15	3117	13,787937	4,234823	12,262498	6,295508
KEA 15	3117	14,999038	**10,67334**	**33,620806**	**16,202877**

Table 1 shows the results for RAKE and KEA on Polish corpus of abstracts. In this validation both of them use Polish lemmatizer. KEA outperforms RAKE. The best F-measure is achieved by *KEA 5,* the highest recall is reported by *KEA 15*.

4.2. PKE Evaluation

We tested PKE with its all possible variants on Polish corpus of abstracts. There were used the same training and validation sets, which were used for RAKE and KEA evaluation.

Table 2 has column *PKE variants,* which define four modes:

- PR unsuper – PKE with pattern-recursive selector, and unsupervised candidate evaluator;
- PR super – PKE with pattern-recursive selector, and supervised candidate evaluator;
- PoS super – PKE with PoS selector, and supervised candidate evaluator;
- PoS unsuper – PKE with PoS selector, and unsupervised candidate evaluator.

Each name in *PKE variants* column ends with a number describing upper limitation of the number of retrieved keywords.

Tab. 2. Results for PKE on Polish abstracts (tbe best results in bold)
Tab. 2. Wyniki algorytmu PKE na zbiorze Polskich abstraktów (najlepsze wyniki pogrubiono)

PKE variants	Abstract no.	mean no. of extracted keywords	precision	recall	F-measure
PR unsuper 5	3117	5	19,108117	20,064681	19,57472
PoS unsuper 5	3117	5	16,689124	17,524592	17,096658
PR super 5	3117	5	**23,298043**	**24,464358**	**23,86696**
PoS super 5	3117	5	21,931344	23,029241	22,466888
PR unsuper 10	3117	9,99615	15,902818	33,384989	21,543478
PoS unsuper 10	3117	9,997754	14,469082	30,380003	19,602217
PR super 10	3117	9,99615	**17,738623**	**37,238917**	**24,030435**
PoS super 10	3117	9,997754	16,920707	35,527557	22,923595
PR unsuper 15	3117	14,928778	13,457116	42,191079	20,405703
PoS unsuper 15	3117	14,954443	12,775406	40,122625	19,380034
PR super 15	3117	14,928778	**14,396235**	**45,135426**	**21,829735**
PoS super 15	3117	14,954443	14,071182	44,192157	21,3457

PR candidate's selector influences the higher precision and recall. Having the same candidate's evaluation method and the same maximal limit of keywords PKE with PR selector outperforms PKE with PoS selector.

Naive Bayes Classifier is a better candidate's evaluation method than unsupervised function *freq(w)^2*. Having the same candidate's selector and the same maximal limit of keywords supervised PKE reports about 4-5% higher recall and 2-4% higher precision than unsupervised one.

The best F-measure was reported by *PR super 10 (24,03%),* and the highest recall is assigned to *PR super 15 (45,13%).*

Comparing RAKE, KEA and PKE it can be very easily noticed that PKE (in all possible modes) is a winner. Only in the one condition KEA is a little bit better (KEA 5 has higher F-measure than PoS unsuper 5). In other cases PKE works much more better, especially all supervised variants surpass KEA and RAKE. The best results are reported by PKE in the supervised mode with the pattern recursive candidate selector (noted as PR super).

This shows that PKE presents relevant improvements in retrieving Polish keywords. Recall is even 12% higher than KEA reports, F-measure notifies increase about 5-7% comparing to KEA metrics.

5. Conclusions

In this study, we present a few methods that are capable of extracting keywords from single Polish documents. We tested different methods on the sets of Polish abstracts of academic papers. We compared the automatic proposed keywords with the authors' keywords and analyzed the results. We designed, implemented and evaluated the algorithm for automatically extracting keywords from a Polish text. Our results show that PKE achieves better results than the other methods (RAKE and KEA - well-known algorithms customized to work with a Polish text). The analysis of candidates selection approaches proved that Pattern Recursive candidate's selector provided better precision and recall measures. The survey related with candidates evaluation methods shows also that Naive Bayes Classifier is a better algorithm than statistic function *freq(w)^2*. PKE's performance is sufficient for providing extraction of keywords from documents in the case where manual keywords assignment is infeasible or insufficient.

In this paper we have evaluated the knowledge-poor methods (not using any external knowledge resources as Wikipedia), which have a limited number of features. Our next goal is to improve the proposed method enhancing it with the vknowledge from external knowledge bases.

6. References

[1] HaCohen-Kerner Y.: Automatic Extraction of Keywords from Ab-stracts. Proceedings of the Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, pp. 843-849, 2003.

[2] Justeson J., Katz S.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, vol.1, pp.9-27, 1995.

[3] Daille B., Gaussier E., Lange J.: Towards automatic extraction of monolingual and bilingual terminology. Proceedings of COLING, pp. 515-521, 1994.

[4] Jones K.: A statistical interpretation of term specififfity and its application in retrieval. Journal of Documentation, pp.11-21, 1972.

[5] Andrade M., Valencia A.: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. BioInformatics, pp.600-607, 1998.

[6] Hulth A.: Improved automatic keyword extraction given more linguistic knowledge. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003.

[7] Witten I.H., Paynter G.W., Frank E., Gutwin C. and Nevill-Manning C.G.: KEA: Practical automatic keyphrase extraction. Working Paper 00/5, Department of Computer Science, The University of Waikato, 2000.

[8] Rose S., Engel D., Cramer N., Cowley W.: Automatic keyword extraction from individual documents. Text Mining: Applications and Theory, pp.19-37, 2010.

otrzymano / received: 06.02.2014
przyjęto do druku / accepted: 01.04.2014 *artykuł recenzowany / revised paper*

INFORMACJE

Wersja elektroniczna miesięcznika PAK

Artykuły opublikowane w PAK po roku 1989 są dostępne w wersji elektronicznej m.in. w bazie artykułów PAK (www.pak.info.pl), w folderze „Archiwum numerów miesięcznika PAK”:

- pełne teksty artykułów z poprzednich lat i streszczenia artykułów najnowszych można pobrać bezpłatnie,
- pełne teksty artykułów z bieżącego roku można otrzymać za opłatą (5 PLN +1,15 PLN VAT).

Informacja redakcji dotycząca artykułów współautorskich

W miesięczniku PAK od numeru 06/2010 w nagłówkach artykułów współautorskich wskazywany jest autor korespondujący (Corresponding Author), tj. ten z którym redakcja prowadzi wszelkie uzgodnienia na etapie przygotowania artykułu do publikacji. Jego nazwisko jest wyróżnione drukiem pogrubionym. Takie oznaczenie nie odnosi się do faktycznego udziału współautora w opracowaniu artykułu. Ponadto w nagłówku artykułu podawane są adresy korespondencyjne wszystkich współautorów.

Wprowadzona procedura wynika z międzynarodowych standardów wydawniczych.

Redakcja