A Language-Independent Approach to Keyphrase Extraction and Evaluation

Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä and Timo Honkela
Adaptive Informatics Research Centre, Helsinki University of Technology
first.last@tkk.fi

Abstract

We present *Likey*, a language-independent keyphrase extraction method based on statistical analysis and the use of a reference corpus. *Likey* has a very light-weight preprocessing phase and no parameters to be tuned. Thus, it is not restricted to any single language or language family. We test *Likey* having exactly the same configuration with 11 European languages. Furthermore, we present an automatic evaluation method based on Wikipedia intra-linking.

1 Introduction

Keyphrase generation is an approach to collect the main topics of a document into a list of phrases. The methods for automatic keyphrase generation can be divided into two groups: keyphrase assignment and keyphrase extraction (Frank et al., 1999). In keyphrase assignment, all potential keyphrases appear in a predefined vocabulary and the task is to classify documents to different keyphrase classes. In keyphrase extraction, keyphrases are supposed to be available in the processed documents themselves, and the aim is to extract these most meaningful words and phrases from the documents.

Most of the traditional methods for keyphrase extraction are highly dependent on the language used and the need for preprocessing is extensive, e.g. including part-of-speech tagging, stemming, and use of stop word lists and other language-dependent filters.

1.1 Related Work

In the statistical keyphrase extraction, many variations for term frequency counts have been proposed in the literature including relative frequencies (Damerau, 1993), collection frequency (Hulth, 2003), term frequency–inverse document frequency (*tf.idf*) (Salton and Buckley, 1988), among others. Additional features to frequency that have been experimented are e.g. relative position of the first occurrence of the term (Frank et al., 1999), importance of the sentence in which the term occurs (HaCohen-Kerner, 2003), and widely studied part-of-speech tag patterns, e.g. Hulth (2003). Matsuo and Ishizuka (2004) present keyword extraction method using word co-occurrence statistical information. Most of the presented methods need a reference corpus or a training corpus to produce keyphrases. The reference corpus acts as a sample of general language, whereas the training corpus is used to tune the parameters of the system.

Statistical keyphrase extraction methods without reference corpora have also been proposed, e.g. (Matsuo and Ishizuka, 2004; Bracewell et al., 2005). The later study is carried out for bilingual corpus.

1.2 Reference Corpora

The reference corpus of natural language processing systems acts as a sample of general language. The corpus should be as large as possible to get sufficiently many examples of language use. In our study, we used the Europarl corpus that consists of transcriptions of European Parliament speeches in eleven European languages, including four Romance languages (Spanish, French, Italian and Portuguese), five Germanic languages

(Danish, German, English, Dutch and Swedish), Finnish and Greek (Koehn, 2005). The number of words in the corpora is between 23 million in Finnish and 38 million in French, while the number of word types differs from 98 thousand in English to 563 thousand in Finnish.

2 The *Likey* Method

We present a keyphrase extraction method *Likey* that is an extension of Damerau's method (Honkela et al., 2007). In Damerau's (1993) method, terms are ranked according to the likelihood ratio and the top m terms are used as index terms. Both single words and bigrams are considered to be terms. *Likey* produces keyphrases using relative ranks of *n*-gram frequencies. It is a simple language-independent method: The only language-specific component is a reference corpus in the corresponding language. *Likey* keyphrases may be single words as well as longer phrases.

The preprocessing phase of *Likey* consists of extraction of the main text body without captions of figures and tables, and removing special characters (except for some hyphens and commas). Numbers are replaced with <NUM> tags.

An integer rank value is assigned to each phrase according to its frequency of occurrence, where the most frequent phrase has rank value one and phrases with the same frequency are assigned the same rank. Rank values rank_a and rank_r are calculated from the text and the reference corpus, respectively, for each phrase. Rank order rank is calculated separately for each phrase length n. Thus we get ranks from unity to $\mathrm{max\_rank}$ for each n. This way n-gram frequencies for $n \geq 2$ are scaled to follow approximately the same distribution as 1-grams in the corpus. The ratio

$$\mathrm{ratio} = \frac{\mathrm{rank}_a}{\mathrm{rank}_r} \tag{1}$$

of ranks is used to compare the phrases.

In highly inflective languages, such as Finnish, and languages with frequent word concatenation, such as German, many of the phrases occurring in the analysed document do not occur in the reference corpus. Thus, their ratio value is related to the maximum rank value, according to Eq. 2,

$$\mathrm{ratio} = \frac{\mathrm{rank}_a}{\mathrm{max\_rank}_r + 1} \tag{2}$$

where $\mathrm{max\_rank}_r$ is the maximum rank in the reference corpus. The ratios are sorted in increasing order and the phrases with the lowest ratios are selected as the extracted keyphrases. Phrases occurring only once in the document cannot be selected as keyphrases.

3 Evaluation

The most straightforward way to evaluate the extracted keyphrases is to first decide which phrases are appropriate to the document and then calculate how many of the extracted keyphrases belong to the appropriate phrases set, e.g. by using precision and recall measures.

There are two widely used approaches for defining the appropriate phrases for a document. The first method is to use human evaluators for rating extracted keyphrases. The other approach is to analyse documents that have author-provided keyword lists. Each document has a list of keyphrases which are easy to accept to be correct. Anyway, automated keyphrase extraction methods are usually poor in predicting author-provided keyphrases since many of the provided phrases do not exist in the document at all but they are sort of superconcepts.

3.1 Multilingual Approach

In our framework, there are keyphrases in 11 languages to be evaluated. Due to many problems related to human evaluation in such a context, we needed a new way of evaluating the results of our language-independent keyphrase extraction method. We took our evaluation data from Wikipedia, a free multilingual online encyclopedia.[1] We present a novel way to use Wikipedia articles in evaluation of a multilingual keyphrase extraction method. Wikipedia corpus has lately been used as a resource for automatic keyword extraction for English (Mihalcea and Csomai, 2007) as well as to many other tasks.

We suppose that those articles which are linked from the article at hand and which link back to the article, are potential keyphrases of the article. For example, a Wikipedia article about some concept may link to its higher-level concept. Likewise, the higher-level concept may list all concepts including to the group.

3.2 Evaluation Data

Finding Wikipedia articles of adequate extent in all the languages is quite challenging, basically due

[1] http://wikipedia.org

to generally quite short articles in Greek, Finnish and Danish. We gathered 10 articles that have sufficient amount of content in each of the 11 Europarl languages. These 110 selected Wikipedia articles were collected in March 2008 and their English names are Beer, Cell (biology), Che Guevara, Leonardo da Vinci, Linux, Paul the Apostle, Sun, Thailand, Vietnam War, and Wolfgang Amadeus Mozart.

The average lengths of articles in Finnish, Dutch and Swedish are below 2 000 words, the lengths of articles in Portuguese, Greek and Danish are around 3 000 words and the rest are between 5 000 and 7 000 words. The normalised lengths would switch the order of the languages slightly.

Among the 67 links extracted from the English Wikipedia article *Cell* include phrases such as *adenosine triphosphate*, *amino acid*, *anabolism*, *archaea*, *bacteria*, *binary fission*, *cell division*, *cell envelope*, *cell membrane*, and *cell nucleus*. The extracted links serve as evaluation keyphrases for the article.

4 Results

In our study, we extracted keyphrases of length $n = 1 \ldots 4$ words. Longer phrases than four words did not occur in the keyphrase list in our preliminary tests. As a baseline, the state-of-the-art keyphrase extraction method *tf.idf* keyphrases were extracted from the same material. *Tf.idf* (Salton and Buckley, 1988) is another simple and non-parameterized language-independent method that can be used for keyphrase extraction. For *tf.idf* we split the Europarl reference corpora in 'documents' of 100 sentences and used the same preprocessing that for *Likey*. To remove uninteresting *tf.idf*-produced phrases like *of the cell*, a *Likey*-like post processing was tried, and it gave slightly better results. Thus the post processing is used for all the reported results of *tf.idf*.

Generally, *Likey* produces longer phrases than *tf.idf*. Each keyphrase list characterises the topic quite well, and most of the extracted keyphrases recur in every language. Both methods extracted a French word *re* that is frequently used in the article as an acronym for *réticulum endoplasmique*. The same word in Dutch is extracted by *tf.idf* in a form *endoplasmatisch reticulum er*.

We compared our *Likey* keyphrase extraction method to the baseline method *tf.idf* by calculating precision and recall measures according to the Wikipedia-based evaluation keyphrases for both methods. We extracted 60 keyphrases from each document for the first evaluation round and the number of keyphrases available in the evaluation keyphrase list for the document for the second evaluation round. Precision and recall values of both *Likey* and *tf.idf* evaluated with Wikipedia intra-links are comparatively low (Table 1) but one has to take into account the nature of the evaluation set with notably varying number of 'correct keyphrases'.

	60 keyphrases		N keyphrases	
Method	Prec.	Recall	Prec.	Recall
Likey	0.1475	0.2470	0.1795	0.1795
tf.idf	0.1225	0.2203	0.1375	0.1375
tf.idf + p	0.1343	0.2341	0.1622	0.1622

Table 1: Average precisions and recalls for *Likey*, *tf.idf* and *tf.idf* with post processing (p). N keyphrases refers to the amount of evaluation keyphrases available for each article.

The obtained precisions and recalls of the first evaluation differed significantly between languages. In Figure 1, the precision and recall of *Likey* and *tf.idf* with post processing for each language is given. Within the 11 European languages, English and German performed best according to the precision (*Likey:* 23.0% and 22.8%, respectively), but not that well according to the recall, where best performed Dutch and Greek (*Likey:* 33.4% and 31.8%, respectively).

5 Conclusions and Discussion

In this paper, we have introduced *Likey*, a statistical keyphrase extraction method that is able to analyse texts independently of the language in question. In the experiments, we have focused on European languages among which Greek and Finnish differ considerably from Romance and Germanic languages. Regardless of these differences, the method gave comparable results for each language.

The method enables independence from the language being analysed. It is possible to extract keyphrases from text in previously unknown language provided that a suitable reference corpus is available. The method includes only lightweight preprocessing, and no auxiliary language-dependent methods such as part-of-speech tagging are required. No particular param-

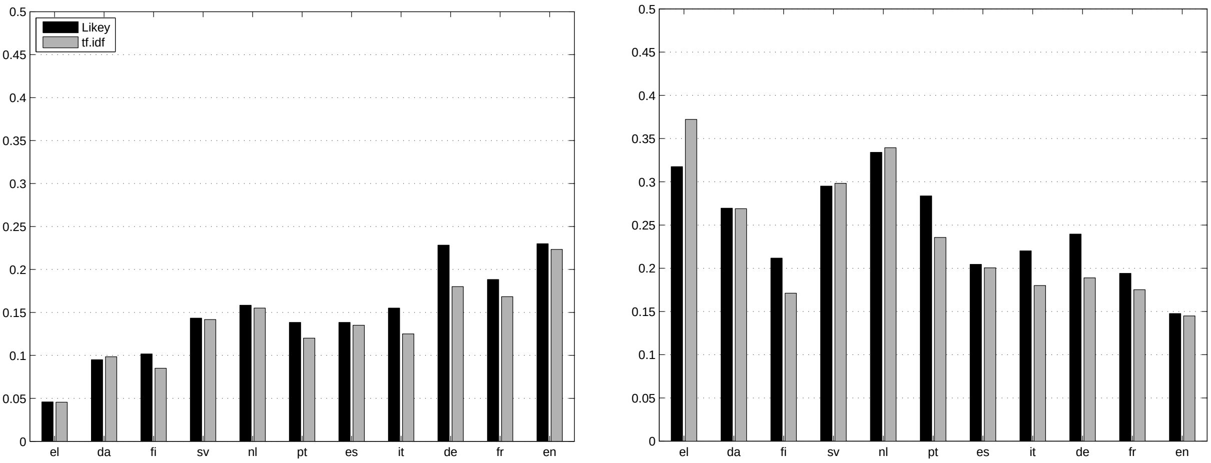

Figure 1: Average precisions (left-hand side) and recalls (right-hand side) of *Likey* and *tf.idf* with post processing for each language. The number of extracted keyphrases is 60.

eter tuning is needed either. A web-based demonstration of *Likey* is available at `http://cog.hut.fi/likeydemo/` as well as more detailed information on the method. The system highlights keyphrases of a document written in one of eleven languages.

Future research includes an extension of *Likey* in which unsupervised detection of morphologically motivated intra-word boundaries (Creutz, 2006) is used. This extension could also handle languages that have no white space between words. We also plan to apply the method within statistical machine translation. A methodological comparison of keyphrase-based dimension reduction and e.g. PCA will also be conducted.

Acknowledgements

This work was supported by the Academy of Finland through the Adaptive Informatics Research Centre that is a part of the Finnish Centre of Excellence Programme. We warmly thank Jaakko J. Väyrynen and Sami Hanhijärvi for their useful comments and ideas.

References

Bracewell, David B., Fuji Ren, and Shingo Kuriowa. 2005. Multilingual single document keyword extraction for information retrieval. In *Proceedings of NLP-KE'05*.

Creutz, Mathias. 2006. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition*. Ph.D. thesis, Helsinki University of Technology.

Damerau, Fred. 1993. Generating and evaluating domain-oriented multi-word terms from text. *Information Processing and Management*, 29(4):433–447.

Frank, Eibe, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *Proceedings of IJCAI'99*, pages 668–673.

HaCohen-Kerner, Yaakov. 2003. Automatic extraction of keywords from abstracts. In Palade, V., R.J. Howlett, and L.C. Jain, editors, *KES 2003, LNAI 2773*, pages 843–849. Springer-Verlag.

Honkela, Timo, Matti Pöllä, Mari-Sanna Paukkeri, Ilari Nieminen, and Jaakko J. Väyrynen. 2007. Terminology extraction based on reference corpora. Technical Report E12, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo. Unpublished.

Hulth, Anette. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 216–223.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

Matsuo, Yutaka and Mitsuru Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *Int'l Journal on Artificial Intelligence Tools*, 13(1):157–169.

Mihalcea, Rada and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA. ACM.

Salton, G. and C. Buckley. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.