Likey: Unsupervised Language-independent Keyphrase Extraction

Mari-Sanna Paukkeri and Timo Honkela
Adaptive Informatics Research Centre
Aalto University School of Science and Technology
P.O. Box 15400, FI-00076 AALTO, Finland
mari-sanna.paukkeri@tkk.fi

Abstract

Likey is an unsupervised statistical approach for keyphrase extraction. The method is language-independent and the only language-dependent component is the reference corpus with which the documents to be analyzed are compared. In this study, we have also used another language-dependent component: an English-specific Porter stemmer as a preprocessing step. In our experiments of keyphrase extraction from scientific articles, the *Likey* method outperforms both supervised and unsupervised baseline methods.

1 Introduction

Keyphrase extraction is a natural language processing task for collecting the main topics of a document into a list of phrases. 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. Keyphrase extraction summarises the content of a document as few phrases and thus provides a quick way to find out what the document is about. Keyphrase extraction is a basic text mining procedure that can be used as a ground for other, more sophisticated text analysis methods. Automatically extracted keyphrases may be used to improve the performance of information retrieval, automatic user model generation, document collection clustering and visualisation, summarisation and question-answering, among others.

This article describes the participation of the *Likey* method in the Task 5 of the SemEval 2010 challenge, automatic keyphrase extraction from scientific articles (Kim et al., 2010).

1.1 Related work

In 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 statistics. An unsupervised keyphrase extraction method by Liu et al. (2009) uses clustering to find exemplar terms that are then used for keyphrase extraction. Most of the presented methods require a reference corpus or a training corpus to produce keyphrases. 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.

2 Data

The data used in this work are from the SemEval 2010 challenge Task 5, automatic keyphrase extraction from scientific articles. The data consist of train, trial, and test data sets. The number of scientific articles and the total number of word tokens in each of the original data sets (before preprocessing) are given in Table 1.

Three sets of "correct" keyphrases are provided for each article in each data set: reader-assigned keyphrases, author-provided keyphrases, and a combination of them. All reader-assigned keyphrases have been extracted manually from the papers whereas some of author-provided

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 162–165,
Uppsala, Sweden, 15-16 July 2010.

Data set	Articles	Word tokens
train	144	1 159 015
trial	40	334 379
test	100	798 049

Table 1: Number of scientific articles and total number of word tokens in the data sets.

keyphrases may not occur in the content. The numbers of correct keyphrases in each data set are shown in Table 2.

Data set	Reader	Author	Combined
train	1 824	559	2 223
trial	526	149	621
test	1 204	387	1 466

Table 2: Number of correct answers in reader, author, and combined answer sets for each data set.

More detailed information on the data set can be found in (Kim et al., 2010).

3 Methods

Likey keyphrase extraction approach comes from the tradition of statistical machine learning (Paukkeri et al., 2008). The method has been developed to be as language-independent as possible. The only language-specific component needed is a corpus in each language. This kind of data is readily available online or from other sources.

Likey selects the words and phrases that best crystallize the meaning of the documents by comparing ranks of frequencies in the documents to those in the reference corpus. The *Likey ratio* (Paukkeri et al., 2008) for each phrase is defined as

$$L(p,d) = \frac{rank_d(p)}{rank_r(p)}, \quad (1)$$

where $rank_d(p)$ is the rank value of phrase p in document d and $rank_r(p)$ is the rank value of phrase p in the reference corpus. The rank values are calculated according to the frequencies of phrases of the same length n. If the phrase p does not exist in the reference corpus, the value of the maximum rank for phrases of length n is used: $rank_r(p) = max\_rank_r(n) + 1$. The *Likey ratio* orders the phrases in a document in such a way that the phrases that have the smallest ratio are the best candidates for being a keyphrase.

As a post-processing step, the phrases of length $n > 1$ face an extra removal process: if one of the words composing the phrase has a rank of less than a threshold ξ in the reference corpus, the phrase is removed from the keyphrase list. This procedure excludes phrases that contain function words such as "of" or "the". As another post-processing step, phrases that are subphrases of those that have occurred earlier on the keyphrase list are removed, excluding e.g. "language model" if "unigram language model" has been already accepted as a keyphrase.

3.1 Reference corpus

Likey needs a reference corpus that is seen as a sample of the general language. In the present study, we use a combination of the English part of Europarl, European Parliament plenary speeches (Koehn, 2005) and the preprocessed training set as the reference corpus. All XML tags of meta information are excluded from the Europarl data. The size of the Europarl corpus is 35 800 000 words after removal of XML tags.

3.2 Preprocessing

The scientific articles are preprocessed by removing all headers including the names and addresses of the authors. Also the reference section is removed from the articles, as well as all tables, figures, equations and citations. Both scientific articles and the Europarl data is lowercased, punctuation is removed (the hyphens surrounded by word characters and apostrophes are kept) and the numbers are changed to `<NUM>` tag.

The data is stemmed with English Porter stemmer implementation provided by the challenge organizers, which differs from our earlier experiments.

3.3 Baselines

We use three baseline methods for keyphrase extraction. The baselines use uni-, bi-, and trigrams as candidates of keyphrases with *tf-idf* weighting scheme. One of the baselines is unsupervised and the other two are supervised approaches. The unsupervised method is to rank the candidates according to their *tf-idf* scores. The supervised methods are *Naïve Bayes (NB)* and *Maximum Entropy (ME)* implementations from WEKA package[1].

[1] `http://www.cs.waikato.ac.nz/~ml/weka/`

4 Experiments

We participated the challenge with *Likey* results of three different parameter settings. The settings are given in Table 3. *Likey-1* has phrases up to 3 words and *Likey-2* and *Likey-3* up to 4 words. The threshold value for postprocessing was selected against the trial set, with $\xi = 100$ performing best. It is used for *Likey-1* and *Likey-2*. Also a bit larger threshold $\xi = 130$ was tried for *Likey-3* to exclude more function words.

Repr.	n	ξ
Likey-1	1–3	100
Likey-2	1–4	100
Likey-3	1–4	130

Table 3: Different parametrizations for *Likey*: n-gram length and threshold value ξ.

An example of the resulting keyphrases extracted by *Likey-1* from the first scientific article in the test set (article C-1) is given in Table 4. Also the corresponding "correct" answers in reader-assigned and author-provided answer sets are shown. The keyphrases are given in stemmed versions. *Likey* keyphrases that can be found in the reader or author answer sets are emphasized.

Likey-1	*uddi registri*, proxi registri, servic discoveri, *grid servic discoveri*, uddi kei, uniqu uddi kei, servic discoveri mechan, distribut hash tabl, *web servic*, *dht*, servic name, web servic discoveri, local proxi registri, local uddi registri, queri multipl registri
Reader	*grid servic discoveri*, uddi, distribut web-servic discoveri architectur, dht base uddi registri hierarchi, deploy issu, bamboo dht code, case-insensit search, queri, longest avail prefix, qo-base servic discoveri, autonom control, *uddi registri*, scalabl issu, soft state
Author	uddi, *dht*, *web servic*, grid comput, md, discoveri

Table 4: Extracted keyphrases by *Likey-1* from article C-1 and the corresponding correct answers in reader and author answer sets.

The example shows clearly that many of the extracted keyphrases contain the same words that can be found in the correct answer sets but the length of the phrases vary and thus they cannot be counted as successfully extracted keyphrases.

The results for the three different *Likey* parametrizations and the three baselines are given in Table 5 for reader-assigned keyphrases and Table 6 for the combined set of reader and author-assigned keyphrases. The evaluation is conducted by calculating precision (P), recall (R) and F-measure (F) for top 5, 10, and 15 keyphrase candidates for each method, using the reader-assigned and author-provided lists as correct answers. The baseline methods are unsupervised *tf-idf* and supervised *Naïve Bayes (NB)* and *Maximum Entropy (ME)*.

Likey-1 performed best in the competition and is thus selected as the official result of *Likey* in the task. Anyway, all *Likey* parametrizations outperform the baselines, *Likey-1* having the best precision 24.60% for top-5 candidates in the reader data set and 29.20% for top-5 candidates in the combined data set. The best F-measure is obtained with *Likey-1* for top-10 candidates for both reader and combined data set: 16.24% and 17.11%, respectively. *Likey* seems to produce the best keyphrases in the beginning of the keyphrase list: for reader-assigned keyphrases the top 5 keyphrase precision for *Likey-1* is 6.8 points better than the best-performing baseline *tf-idf* and the corresponding F-measure is 4.0 points better. For the combined set, the numbers are 7.2 and 3.7 points, respectively. The difference decreases for the larger keyphrase sets.

5 Conclusions and discussion

This article describes our submission to SemEval 2010 Task 5, keyphrase extraction from scientific articles. Our unsupervised and language-independent method *Likey* uses reference corpus and is able to outperform both the unsupervised and supervised baseline methods. The best results are obtained with the top-5 keyphrases: precision of 24.60% with reader-assigned keyphrases and 29.20% with the combination of reader-assigned and author-provided keyphrases.

There are some keyphrases in the answer sets that our method does not find: due to the comparatively large threshold value ξ many phrases that contain function words, e.g. "of", cannot be found. We also extract keyphrases of maximum length of three or four words and thus cannot find keyphrases longer than that. The next step of this research would be to take these problems into account.

Method	Top 5 candidates			Top 10 candidates			Top 15 candidates		
	P %	R %	F %	P %	R %	F %	P %	R %	F %
Likey-1	**24.60**	10.22	14.44	17.90	14.87	**16.24**	13.80	**17.19**	15.31
Likey-2	23.80	9.88	13.96	16.90	14.04	15.34	13.40	16.69	14.87
Likey-3	23.40	9.72	13.73	16.80	13.95	15.24	13.73	17.11	15.23
tf-idf	17.80	7.39	10.44	13.90	11.54	12.61	11.60	14.45	12.87
NB	16.80	6.98	9.86	13.30	11.05	12.07	11.40	14.20	12.65
ME	16.80	6.98	9.86	13.30	11.05	12.07	11.40	14.20	12.65

Table 5: Results for *Likey* and the baselines for the reader data set. The best precision (P), recall (R) and F-measure (F) are highlighted.

Method	Top 5 candidates			Top 10 candidates			Top 15 candidates		
	P %	R %	F %	P %	R %	F %	P %	R %	F %
Likey-1	**29.20**	9.96	14.85	21.10	14.39	**17.11**	16.33	**16.71**	16.52
Likey-2	28.40	9.69	14.45	19.90	13.57	16.14	15.73	16.10	15.91
Likey-3	28.00	9.55	14.24	19.60	13.37	15.90	16.07	16.44	16.25
tf-idf	22.00	7.50	11.19	17.70	12.07	14.35	14.93	15.28	15.10
NB	21.40	7.30	10.89	17.30	11.80	14.03	14.53	14.87	14.70
ME	21.40	7.30	10.89	17.30	11.80	14.03	14.53	14.87	14.70

Table 6: Results for *Likey* and the baselines for the combined (reader+author) data set. The best precision (P), recall (R) and F-measure (F) are highlighted.

Acknowledgements

This work was supported by the Finnish Graduate School in Language Studies (Langnet) funded by Ministry of Education of Finland.

References

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

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

Eibe Frank, 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.

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

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

Su Nam Kim, Alyona Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010)*. to appear.

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

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 257–266, Singapore, August. Association for Computational Linguistics.

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

Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä, and Timo Honkela. 2008. A language-independent approach to keyphrase extraction and evaluation. In *Coling 2008: Companion volume: Posters*, pages 83–86, Manchester, UK, August. Coling 2008 Organizing Committee.

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