UNPMC: Naïve Approach to Extract Keyphrases from Scientific Articles

Jungyeul Park
LINA,
Université de Nantes
Nantes, France
jungyeul.park
@univ-nantes.fr

Jong Gun Lee
LIP6-CNRS,
UPMC (Paris 6)
Paris, France
jonggun.lee
@lip6.fr

Béatrice Daille
LINA,
Université de Nantes
Nantes, France
beatrice.daille
@univ-nantes.fr

Abstract

We describe our method for extracting keyphrases from scientific articles which we participate in the shared task of SemEval-2 Evaluation Exercise. Even though general-purpose term extractors along with linguistically-motivated analysis allow us to extract elaborated morpho-syntactic variation forms of terms, a naïve statistic approach proposed in this paper is very simple and quite efficient for extracting keyphrases especially from well-structured scientific articles. Based on the characteristics of keyphrases with section information, we obtain 18.34% for f-measure using top 15 candidates. We also show further improvement without any complications and we discuss this at the end of the paper.

1 Introduction[1]

Key phrases are a set of words to capture the main topic of the document. Since key phrases contain the substance of the document, they are used in the large spectrum of areas; from applications which explicitly use key phrases such as automatic indexing, documents classification and search engine optimization in information retrieval, to applications which implicitly use key phrases such as summarization and question-answering systems. During the last decade, many previous works have dealt with the various methods for automatically extracting key phrases (e.g., Frank et al., 1999; Barker and Corrnacchia, 2000; Turney, 2003; Medelyan and Witten, 2006; Nguyen and Kan, 2007; Wan and Xiao, 2008).

The task of extracting key phrases would be considered as a subtask of extracting terminology if key phrases are a kind of terms. Typical approaches for automatically extracting terms use linguistic preprocessing which involves morpho-syntactic analysis such as part-of-speech tagging and phrase chunking, and statistical postprocessing such as log likelihood which compares the term frequencies in a document against their expected frequencies derived in a bigger text. Besides, extracting terms prefers syntactically plausible noun phrases (NPs) which are mainly multi-words terms. Kim and Kan (2009) report that most of key phrases are often simple words than less often compound words[2].

The task for extracting key phrases tend to include analyzing the document structure. Especially, extracting key phrases from well-structured scientific articles should consider cross-section information (Nguyen and Kan, 2007). This information has been explored to assess the suitability of features during learning in Kim and Kan (2009).

Extracting key phrases, however, is more than to extracting terminology or analyzing the document structure. While terms are words which appear in specific contexts and analyse concept structures in *domains* of human activity, key phrases are words that capture the key idea of *documents*. In addition, while terms usually occur in the given document more often than we would expect to occur, key phrases do not necessarily occur frequently or key phrases do not occur at all in the document. Consequently, the task for extracting key phrases should not be considered as the subtask of extracting terminology and we are not able to directly apply general-purpose term extractors to extract key phrases.

In this paper, we describe our method for "Automatic Keyphrase Extraction from Scientific Ar-

[1] UNPMC means the collaborative team from Laboratoire d'Informatique de Nantes Atlantique of the Université de Nantes and Laboratoire d'Informatique de Paris 6 of the Université Pierre et Marie Curie.

[2] In training data, only 23.4% of keyphrases, however, are single words.

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 178–181,
Uppsala, Sweden, 15-16 July 2010.

ticles", the shared task of SemEval-2 Evaluation Exercise which we participated in. Although term extractors along with linguistically-motivated analysis allow us to extract even elaborated morpho-syntactic variation forms of terms, the naïve statistic approach proposed in this paper is very simple and quite efficient for extracting keyphrases especially from well-structured scientific articles. In a nutshell, our method is based on empirical rules without any linguistically-motivated preprocessing. Empirical rules are obtained from the analysis of the characteristics of keyphrases by observing training data.

The remaining of this paper is organized as follows: Section 2 explains the characteristics of keyphrases in scientific articles. Section 3 and 4 detail our naïve statistic approach and experiment, respectively. We conclude this paper and discuss a further improvement in Section 6.

2 Characteristics of Keyphrases in Scientific Articles

In this section, we investigate the characteristics of keyphrases in training data. Table 1 shows statistics of training data. In Table 1, `D-author` means the keyphrases assigned by authors, `D-reader` the keyphrases assigned by readers, and `D-combined` the combined keyphrases assigned by both of authors and readers.

	# of papers (p)	# of key phrases (k)	k / p
`D-author`	144	563	3.91
`D-reader`	144	1,865	12.95
`D-combined`	144	2,265	15.73

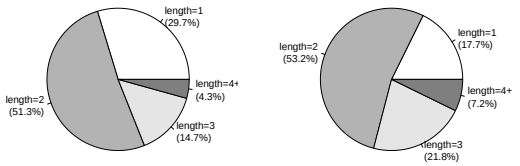
Table 1: Statistics of training data

2.1 Word length of keyphrases

We measure the distribution of word length of key phrases in training data and present it in Figure 1. Over half of key phrases are two-word key phrases in both author- and reader-assigned key phrases. Differently with Kim and Kan (2009) which they reported that most of key phrases are often simple words than less often compound words, only 29.7% and 17.7% of key phrases are one-word key phrases. There are also more than four-word key phrases which hold 4.3% and 7.2% of author and reader assigned key phrases, respectively.

2.2 Occurrences of keyphrases

In which section do keyphrases occur frequently? To answer this question, we count the number of occurrences of keyphrases of each section. Due to the variation of the naming of the section, we divide sections into title and abstract, introduction, conclusion, and the rest including references. Table 2 and 3 show the number of occurrences and the accumulative number of unique occurrences of keyphrases in each section, respectively. We also show the accumulative number of words in each section in Table 4. Including the rest sections exponentially diminishes the ratio of the number of gold keyphrases to the number of candidate keyphrases. Note that m words produce $\sum_{i=0}^{n-1}(m-i)$ candidate keyphrases for up to n-word keyphrases by supposing that candidate keyphrases are simple n-word terms.



(a) `D-author` (b) `D-reader`

Figure 1: Word length of keyphrases in training data

Note also that both author- and reader-assigned keyphrases hold only 75.49% and 89.44%, respectively. Even some keyphrases are different with surface forms in the document and our naïve method with no linguistic intervention is not able to recognize them. For example, one of reader-assigned keyphrases *distributed real-time embedded system* for C-41 actually appears as *distributed real-time and embedded (DRE) systems*.

	`D-author`	`D-reader`
Title and Abstract	277	802
Introduction	215	491
Conclusion	313	982
Other	387	1,210

Table 2: Number of occurrences of keyphrases in each section

	`D-author`	`D-reader`
Total	563 (100.0%)	1,865 (100.0%)
Title and Abstract	277 (49.20%)	802 (43.00%)
'+' Introduction	317 (56.30%)	937 (50.24%)
'+' Conclusion	367 (65.19%)	1,311 (70.29%)
'+' Other	425 (75.49%)	1,668 (89.44%)

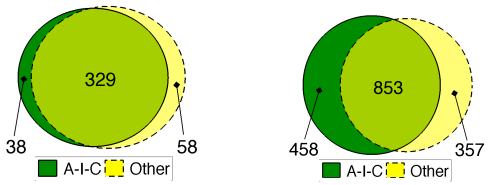
Table 3: Accumulative number of unique occurrences of keyphrases in each section

	# words (W)	# gold (G)	G/W
Title and Abstract	28435	802	0.0282
'+' Introduction	72729	937	0.0128
'+' Conclusion	178473	1311	0.0073
'+' Other	948007	1668	0.0018

Table 4: Number of words in training data and gold data (D-reader)

2.3 Coincidence of keyphrases

Figure 2 shows the coincidence of keyphrases[3]. Almost half of keyphrases (58.44% and 45.74% for author- and reader-assigned keyphrases, respectively) occur coincidentally in keysections and the rest sections. Keysections hold 65.19% and 70.29% of keyphrases and the rest sections besides keysections hold 68.74% and 64.88% of whole keyphrases. Note that the rest sections occupy over 70% of the document on the average.



(a) D-author (b) D-reader

Figure 2: Coincidence of keyphrases

3 Methodology

From training data, we observe and decide the followings:

- More than four-word keyphrases hold only 4.3% and 7.2% of whole keyphrases. We decide that our approach limits the word length as three for extracting keyphrases. Thus we extract only up to three-word keyphrases. This choice might lead the performance degradation of our method because we explicitly exclude more than four-word keyphrases.

- Keysections hold 65.19% and 70.29% of keyphrases. We decide that our approach limits keysections from which we extract keyphrases. Including the rest sections may improve recall, but probably diminish precision since the rest sections occupy over 70% of the document.

- Almost half of keyphrases occur coincidentally in keysections and the rest sections. We decide that our approach limits coincident keyphrases in both of them. This decision is made empirically and improve precision.

The following procedure explains and details our approach for extracting keyphrases.

- Extract up to three-word terms from keysections as candidate keyphrases.
- Filter them out if they contain one or more of stop words or non-content-containing words (see Table 5 for non-content-containing words).
- Count the number of occurrences of extracted terms from each keysection.
- Check the coincidence whether candidate keyphrases occurs in more than two keysections. If so, we assign weight.
- Calculate a score for candidate keyphrases and list them by order of the score.

4 Experiment results

This section shows the experiment results with training and test data.

4.1 Training data

To optimize our results, we use various thresholds for the number of *n*-word keyphrases and weight.

We try to find the $(i : j : k)$ pattern which means i one-word, j two-word, and K three-word keyphrases to produce the best results. We also try to find the threshold for weight d to calculate the score as follows: if keyphrases appear in more than two keysections, $score = d * \#\ of\ total\ occurences$, otherwise $score = \#\ of\ total\ occurences$. Table 6 shows our best results for training data where $(i : j : k) = (3 : 9 : 3)$ and $d = 2$. Empirically, we found these thresholds from training data by iterating several possibilities[4].

4.2 Test data

Table 7 shows our test data results published by organizers of the shared task of SemEval-2 Evaluation Exercise.

[3] We denote title and abstract as A, introduction as I, conclusion as C, and the rest sections including references as Other.

[4] These thresholds will be more examined in future work.

Type	Examples
Noun	section, abstract, introduction, conclusion, reference, future work, figure, paper, result, laboratory, university
Verb	present, how, introduce, become, improve, find, help, improve, consider, call, yield, allow, give, assume
Adverb	always, formally, necessarily, successfully, previously, usually,mainly, final, essentially, ultinately, commonly, severely, significantly, dramatically, clearly, still, well, who, whose, whom, which, whether, therefore,
Other POSs	that, this, those, these, many, several, more, over, less, behind, above, below, each, few, different, under, both, within, through, prior, various, better, following, between, possible, via, before,even, such, if, new, show, important, simple, good, tranditional, current, varying, necessary, previous, clear

Table 5: Example of (heuristically obtained) non-content-containing terms

AUTHOR.STEM.FINAL				
# Gold: 559	Match	Precision	Recall	F-score
Top 05	43	5.97%	7.69%	6.72%
Top 10	101	7.01%	18.07%	10.10%
Top 15	139	6.44%	24.87%	10.23%

READER.STEM.FINAL				
# Gold: 1824	Match	Precision	Recall	F-score
Top 05	118	16.39%	6.47%	9.28%
Top 10	249	17.29%	13.65%	15.26%
Top 15	361	16.71%	19.79%	18.12%

COMBINED.STEM.FINAL				
# Gold: 2223	Match	Precision	Recall	F-score
Top 05	143	19.86%	6.43%	9.71%
Top 10	309	21.46%	13.90%	16.87%
Top 15	441	20.42%	19.84%	20.13%

Table 6: Training data results

READER.STEM.FINAL			
# Gold: 1204	Precision	Recall	Fscore
Top 05	13.80%	5.73%	8.10%
Top 10	15.10%	12.54%	13.70%
Top 15	14.47%	18.02%	16.05%

COMBINED.STEM.FINAL			
# Gold: 1466	Precision	Recall	Fscore
Top 05	18.00%	6.14%	9.16%
Top 10	19.00%	12.96%	15.41%
Top 15	18.13%	18.55%	18.34%

Table 7: Test data results

5 Conclusion and Discussion

In this paper, we described our simple method for extracting keyphrases from scientific articles which we participate in the shared task of SemEval-2 Evaluation Exercise. The naïve approach was proposed. This approach turned out very simple and quite efficient for extracting keyphrases from well-structured scientific articles. Based on learning the distribution of keyphrases with section information, we obtain 18.34% for f-measure using top 15 candidates.

Our naïve approach still has much room for improvement. For example, we are able to improve the result for same test data up to 20.71% and 25.55% for f-measure using top 15 candidates simply by adding the rest sections and normalizing the number of occurrences of terms from each section[5].

[5]The result is not improved only by adding the rest sections.

Moreover, our *n*-word terms based extraction can be benefited by linguistic preprocessing such as normalizing surface forms. Handcrafted regular expression rules along with part-of-speech tagging and phrase chunking would be also introduced to improve candidate selection. We have not explored thoroughly feature engineering, neither. For example, more fine-grained section information and weight re-assignment might help filter out irrelevant candidates. We leave these possibilities for future work.

References

Ken Barker and Nadia Cornacchia. 2000. Using noun phrase heads to extract document keyphrases. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 40-52. May 14-17, 2000. Montréal, Quebec, Canada.

Eibe Frank , Gordon W. Paynter , Ian H. Witten , Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-Specific Keyphrase Extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 668-673. July 31-August 6, 1999. Stockholm, Sweden.

Su Nam Kim and Min-Yen Kan. 2009. Re-examining Automatic Keyphrase Extraction Approaches in Scientific Articles. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009), ACL-IJCNLP 2009*, pages 9-12. August 6, 2009. Singapore.

Olena Medelyan and Ian H. Witten. 2006. Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 296-297. June 11-15, 2006. Chapel Hill, NC, USA.

Thuy Dung Nguyen and Min-Yen Kan. 2007. Key phrase Extraction in Scientific Publications. *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 317-326. Springer Berlin, Heidelberg.

Peter D. Turney. 2003. Coherent keyphrase extraction via Web mining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 434-439. August 9-15, 2003. Acapulco, Mexico.

Xiaojun Wan and Jianguo Xiao. 2008. CollabRank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 969-976. 18-22 August, 2008. Manchester, UK.