Unsupervised Keyphrase Extraction: Introducing New Kinds of Words to Keyphrases

Tho Thi Ngoc Le, Minh Le Nguyen, and Akira Shimazu

Japan Advanced Institute of Science and Technology
{tho.le, nguyenml, shimazu}@jaist.ac.jp

Abstract. Current studies often extract keyphrases by collecting adjacent important adjectives and nouns. However, the statistics on four public corpora shows that about 15% of keyphrases contain other kinds of words. Even so, incorporating such kinds of words to the noun phrase patterns is not a solution to improve the extraction performance. In this work, we propose a solution to improve the extraction performance by involving new kinds of words to keyphrases. We have experimented on four public corpora to demonstrate that our proposal improve the performance of keyphrase extraction and new kinds of words are introduced to keyphrases. In addition, our proposal is also superior to the current unsupervised keyphrase extraction approaches.

Keywords: automatic keyphrase extraction, syntactic structure, parse tree, unsupervised approach

1 Introduction

Keyphrases are single-token or multi-token expressions that provide the essential information of a sentence or document. Automatic keyphrase extraction plays an important role in many applications of natural language processing (NLP). Many approaches have been proposed for extracting keyphrases automatically. Those approaches contain two common steps: 1) collecting as many tokens as possible for candidates which benefit keyphrase extraction; and 2) combining adjacent candidate tokens to obtain keyphrases using a fixed pattern of adjectives and nouns. Up to now, candidates for keyphrases are adjectives and nouns which are collected by many methods: applying linguistic knowledge (e.g. syntactic features like POS tags, NP chunks) and statistics (e.g. term frequency, inverse document frequency, n-grams) [13, 1, 4]; applying graph-based ranking technique [10]; or applying clustering technique [9, 7]. An overview of approaches for automatic keyphrase extraction can be found in a survey by Hasan and Ng [3].

Since previous research applies a fixed pattern to extract keyphrases, i.e. adjacent adjectives and nouns, candidate tokens are restricted to a set of pre-specified words of only adjectives and nouns. This restriction causes a consequence that other kinds of words cannot be selected as candidates, and therefore never appear in keyphrases. For that reason, extraction still vary in a certain performance. Practically, not all of keyphrases are composed of only adjectives and nouns. Indeed, when shedding a light on the patterns of keyphrases in four public corpora, we found that there are approximately 15% of keyphrases contain words other than adjectives and nouns. Unfortunately, the extraction performance is possibly decreased when we involve more kinds of

words to the pattern of keyphrases since the extracted keyphrases are composed into incorrect grammar phrases. However, since other kinds of words do appear in keyphrases, there should be an investigation for a novel approach which tackling new words to keyphrase extraction.

In this work, we motivate to introduce words other than adjectives and nouns, which benefit the extraction performance, to keyphrases. We propose a novel approach to extract keyphrases by collecting noun phrases (NPs) as candidate keyphrases using syntactic information, i.e. chunks and constituent syntactic parse trees. Hence, the well-formedness of keyphrases are ensured by noun phrases from chunks and parse trees. In addition, words other than adjectives and nouns are also considered to keyphrases pattern if they appear in noun phrase candidates. We experimented keyphrase extraction on four public corpora and achieved very competitive performance. Compared to extraction using patterns and the whole chunks, our proposal takes advantage in performance while reserving the well-formedness of keyphrases and involving more kinds of words. We achieve better performance than the state-of-the-art on three corpora while we are still behind a supervised approach, which employs many features for machine learning. Therefore, we are able to conclude that our proposed approach is a competitive approach for unsupervised keyphrase extraction.

2 Corpora and Keyphrase Analysis

We consider four public corpora, namely DUC-2001 [12, 14], Inspec [4], NUS [11], and SemEval-2010 [5], which are used for evaluating the extraction performance in previous studies. Some characteristics of these corpora have been analyzed in previous studies [2, 5]. In this work, we examine two other characteristics in concern of our work and show them in Table 1. The characteristics in our concern are: the percentage of one-word keyphrases and the percentage of the types of keyphrase patterns. Each corpus has a different percentage of one-word keyphrases and the percentage of one-word

Table 1: The characteristics of four public corpora of keyphrase extraction.

	Corpora			
	DUC-2001	**Inspec**	**NUS**	**SemEval-2010**
Type	News articles	Paper abstracts	Paper abstracts	Paper abstracts
# Documents for test	308	500	211	100
# Keys	2,484	4,913	2,327	1,482
# One-word keys	431 (17.4%)	659 (13.4%)	610 (26.2%)	309 (20.9%)
# Keys (adj+noun)	2,298 (92.5%)	4,221 (85.9%)	1,903 (81.8%)	(84.5%)
# Keys (w. participles)	53 (2.1%)	383 (7.8%)	206 (8.9%)	(7.2%)
# Keys (other patterns)	133 (5.4%)	309 (6.3%)	218 (9.4%)	(8.3%)
# Exist. keys in text	2,462 (99.1%)	3,826 (77.9%)	2,200 (94.5%)	(89.5%)
# Exist. keys (adj+noun)	2,277 (91.7%)	3,338 (68%)	1,837 (78.9%)	(80%)
# Exist. keys (w. participles)	53 (2.1%)	287 (5.8%)	178 (7.6%)	(3.2%)
# Exist. keys (other patterns)	132 (5.3%)	201 (4.1%)	185 (8%)	(6.3%)

keyphrases on four corpora is 19.5% on average. Since this is a significant percentage, a certain percentage of one-word keyphrases should be specified when extracting keyphrases.

When analyzing the patterns of keyphrases, we observe that not only adjectives and nouns appear in keyphrases, but other types of words also appear, such as: participles (*watermarking*, *ordering criteria*, *synthesized data*); adverbs (*highly nonlinear rule-based models*, *visually impaired people*, *partially ordered set*); cardinal numbers (*four main design patterns*, *category 5 hurricane*, *type II diabetes*).

The percentage of keyphrases for each type of keyphrase pattern is showed next to the number of keyphrases in Table 1. Note that, keyphrases in test set of SemEval-2010 are provided as stemmed words, we examine the characteristic of keyphrases in training set instead. As shown in Table 1, the percentage of keyphrases which follow the patterns of adjectives and nouns is 86.2% on average. When looking closely to the annotated keyphrases, about 90% keyphrases actually exist in the text. The percentage of existing keyphrases which follow the patterns of adjectives and nouns is 79.6% on average. Consequently, the highest recall of extraction performance is less than 80% when involving only adjectives and nouns.

Among other types of words in keyphrases, verbs in forms of present and past participles have a considerable contribution to keyphrases. Therefore, in the following section, we examine whether involving participles as candidates in keyphrase patterns improves the extraction performance.

3 Analysis on Extracting Candidates using Noun Phrase Patterns

This section describes the extracting process using patterns of noun phrases. Since approximately 86% of keyphrases are combinations of adjectives and nouns, to collect noun phrases, previous works often use a pattern to collapse adjacent adjectives and nouns. When examine the keyphrases pattern, we recognize that adjectives in forms of comparative and superlative also appear, e.g. *lower net income* and *nearest parent model*. Hence, the pattern for noun phrases is modified as the following regular expression

$$(JJ|JJR|JJS)*(NN|NNS|NNP|NNPS)+$$

As analyzed in Section 2, on average 6.5% of verbs in forms of present and past participles play the roles as adjectives and nouns in keyphrases. Hence, we involve them to the pattern of noun phrases and introduce another pattern for noun phrases to examine whether including such participles improves extraction performance as following

$$(JJ|JJR|JJS|VBG|VBN)*(NN|NNS|NNP|NNPS|VBG)+$$

We assign weights to candidates and extract keyphrases as the work by Le et al. [6]. Experiments are run on four public corpora described in Section 2 and followed the evaluation criteria in SemEval-2010. We extract up to 15 highest weighted keyphrases for each document including one single keyphrases and 14 compound keyphrases. This technique is compared to a baseline [2, 5], henceforth referred as *TFIDF n-grams* for convenience. In TFIDF n-grams, the top weighted *n*-grams of adjectives and nouns are

Table 2: The performance of keyphrase extraction using patterns.

	DUC-2001			Inspec			NUS			SemEval-2010		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Pattern(adj+noun)	21.3	**39.6**	**27.7**	35.4	44.9	39.6	15.7	20.2	**17.7**	21.0	20.7	**20.8**
Pattern(+participle)	20.8	38.8	27.1	34.0	44.6	38.6	14.9	19.6	16.9	19.4	19.4	19.4
TFIDF *n*-grams	-	-	27.0	-	-	36.3	-	-	6.6	14.9	15.3	15.1

extracted as keyphrases where the weight of a candidate is calculated by summing its constituent unigrams.

The performance of the proposed technique is presented in Table 2. Our technique achieves better performance than the TFIDF n-grams baseline for all corpora. Henceforth, we use these results as new baseline for keyphrase extraction in this work. When adding verbs in forms of present and past participles to the pattern of keyphrases, the performance decreases. The reason is that the participles which modify the meaning of noun phrases are confused with the verbs of sentences. For an example, considering the sentence *"Previous research has **indicated differing levels** of importance of perceived ease of use relative to other factors,"* the phrase "*indicated differing levels*" is wrongly extracted as a keyphrase since it satisfies the pattern of keyphrases. In fact, the participle "*indicated*" does not modify the meaning of noun phrase "*differing levels*" but it is a conjugation of verb in present participle tense.

Based on experimental results, we conclude that the performance of keyphrase extraction is not improved when involving present and past participles into noun phrase patterns. However, as keyphrases contain such parts-of-speech of words, an approach should be investigated to capture all possible words to keyphrases.

4 Extracting Candidates using Syntactic Information

This section introduces a novel technique to improve the performance of keyphrase extraction by exploiting the syntactic information with two levels: *shallow (chunks)* and *deep (constituent parse trees)*. Our proposal to extract keyphrases using syntactic information is outlined as follows:

1. Parse all sentences in document for syntactic information;
2. Collect noun phrases as candidates;
3. Post-process candidates to make sure they are well-formed;
4. Assign weights to candidates to indicate their importance;
5. Collect the top weighted candidates as the keyphrases.

Weights of candidates are also computed as the way in work by Le et al. [6]. The post-processing step is to ensure that candidates are well-formed. This step removes the unnecessary tokens from the beginning and ending of candidates. Two ways are introduced to remove unnecessary tokens to ensure that:

– A Candidate begins with a token whose POS tag is JJ, JJR, JJS, NN, NNS, NNP, or NNPS; and ends with a token whose POS tag is NN, NNS, NNP, or NNPS.

– A Candidate begins with a token whose POS tag is JJ, JJR, JJS, NN, NNS, NNP, NNPS, VBG, or VBN; and ends with a token whose POS tag is NN, NNS, NNP, NNPS, or VBG.

In other words, we consider only adjectives and nouns in the first way while involving participles to the candidates in the second way.

Experiments and Evaluations

Illinois Chunker and Stanford CoreNLP tools are respectively employed to parse sentences into chunks and parse trees. Then, noun phrases are extracted using these syntactic information. After that, each noun phrase is post-processed to eliminate the punctuation, conjunctions and unnecessary tokens. In post-processing, each noun phrase is split at the position of punctuation or conjunctions (if any).

The experiments of our proposal are also run on four public corpora. The extraction performance using syntactic information is shown in Table 3 in comparison to two baselines: Pattern baseline (ref. Section 3) and Previous best F1 baseline. For DUC-2001 and NUS corpora, TFDIF *n*-grams yields the state-of-the-art performance [2]. For Inspec corpus, clustering approach [7] achieves highest F1-score. For SemEval-2010 corpus, HUMB [8], a supervised system, obtains the best performance.

The results in Table 3 show that, in all corpora, our proposed approach beats the performance of Pattern baseline. In most of cases, the precision and recall scores are higher. When comparing to the previous best F1 scores, our proposal achieves the best performance on three corpora: DUC-2001, Inspec and NUS. On SemEval-2010 corpus, our approach still behind HUMB because this is a supervised method which exploits many features for machine learning: structure of the article, lexical cohesion of a sequence of words, TFIDF scores, and the frequency of the keyword in the global corpus.

When using the syntactic information for extracting candidates, we found that the recall is generally higher if participles are taken into account. In addition, other kinds of words, e.g. cardinal numbers, which occur in the middle of keyphrases are also included. For examples, keyphrases *modulo 2 residue class*, *category 5 hurricane* and *type II diabetes* are extracted by using syntactic information no matter participles are tackled or not. Even though words other than adjectives and nouns are involved, the syntactic information keeps the well-formedness of keyphrases. Therefore, both the recall and precision are improved.

Table 3: The performance of keyphrase extraction using syntactic information.

	DUC-2001			**Inspec**			**NUS**			**SemEval-2010**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Chunk(adj+noun)	21.3	39.6	27.7	38.6	45.3	**41.7**	17.5	21.7	**19.4**	22.7	21.5	22.1
Chunk(+participle)	21.4	39.8	**27.9**	38.1	46.1	**41.7**	17.2	21.8	19.2	23.4	22.7	**23.1**
Parse tree(adj+noun)	21.1	39.3	27.5	38.4	44.7	41.3	17.0	20.9	18.8	22.5	21.3	21.9
Parse tree(+participle)	21.1	39.3	27.5	37.6	44.8	40.9	16.8	21.0	18.7	21.9	21.1	21.5
Pattern(adj+noun)	21.3	39.6	27.7	35.4	44.9	39.6	15.7	20.2	17.7	21.0	20.7	20.8
Previous best F1	-	-	27.0	-	-	40.6	-	-	6.6	27.2	27.8	**27.5**

5 Conclusions

We have demonstrated that keyphrases are not consistently the combination of adjectives and nouns. There are roughly 15% of keyphrases including other kinds of words such as participles, comparative/superlative adjectives and cardinal numbers. We believe that participles should be considered in keyphrase extraction since there is a recognizable percentage of keyphrases containing participles (6.5%). To improve the extraction performance and to take into account new kinds of words in keyphrases, we proposed to incorporate the syntactic information when extracting noun phrases as keyphrase candidates. We show the experimental results on four public corpora, in which performance has been improved and new kinds of words has been also introduced to the keyphrases.

Acknowledgments. This work was partly supported by JSPS KAKENHI Grant Number JP15K12094.

References

1. Frank E., Paynter G.W., Witten I.H., Gutwin C., and Nevill-Manning C.G.: Domain-Specific Keyphrase Extraction. In: Proc. IJCAI'99, pp. 668–673 (1999)
2. Hasan K.S. and Ng V.: Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-art. In: Proc. COLING'10: Posters, pp. 365–373 (2010)
3. Hasan K.S. and Ng V.: Automatic Keyphrase Extraction: A Survey. In: Proc. ACL'14, pp. 1262–1273 (2014)
4. Hulth A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proc. EMNLP'03, pp. 216–223 (2003)
5. Kim S.N., Medelyan O., Kan M.Y., and Baldwin T.: SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In: Proc. SemEval'10 , pp. 21–26 (2010)
6. Le T.T.N., Nguyen M.L., and Shimazu A.: Unsupervised Keyword Extraction for Japanese Legal Documents. In: Proc. JURIX'13, pp. 97–106 (2013)
7. Liu Z., Li P., Zheng Y., and Sun M.: Clustering to find exemplar terms for keyphrase extraction. In: Proc. EMNLP'09, pp. 257–266 (2009)
8. Lopez P. and Romary L.: HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. In: Proc. SemEval'10, pp. 248–251 (2010)
9. Matsuo Y. and Ishizuka M.: Keyword Extraction From A Single Document Using Word Co-Occurrence Statistical Information. J. Artificial Intelligence Tools, 13:1, pp. 157–169 (2004)
10. Mihalcea R. and Tarau P.: TextRank: Bringing Order into Texts. In: Proc. EMNLP'04, pp. 404–411 (2004)
11. Nguyen T.D. and Kan M.Y.: Keyphrase Extraction in Scientific Publications. In: ICADL'07, pp. 317–326 (2007)
12. Over P.: Introduction to DUC-2001: an Intrinsic Evaluation of Generic News Text Summarization Systems. In: Proc. International 2001 Document Understanding Conference (2001)
13. Turney P.D.: Learning Algorithms for Keyphrase Extraction. J. Information Retrieval, 2:4, pp. 303–336 (2000)
14. Wan X. and Xiao J.: Single document keyphrase extraction using neighborhood knowledge. In: Proc. AAAI'08, vol. 2, pp. 855–860 (2008)