2017

Extended list of stop words: Does it work for keyphrase extraction from short texts?

Svetlana Popova

Gabriella Skitalinskaya

Extended List of Stop Words: Does It Work for Keyphrase Extraction from Short Texts?

Svetlana Popova
Saint-Petersburg State University, Saint-Petersburg, Russia
ITMO University, Saint-Petersburg, Russia
Email: svp@list.ru

Gabriella Skitalinskaya
MIPT (State University), Moscow, Russia
Institute of Technology Tallaght, Dublin, Ireland
Email: gabriellasky@icloud.com

***Abstract*— In this paper we study the problem of key phrase extraction from short texts written in Russian. As texts we consider messages posted on Internet car forums related to the purchase or repair of cars. The main assumption made is: the construction of lists of stop words for key phrase extraction can be effective if performed on the basis of a small, expert-marked collection. The results show that even a small number of texts marked by an expert can be enough to build an extended list of stop words. Extracted stop words allow to improve the quality of the key phrase extraction algorithm. Prior, we used a similar approach for key phrase extraction from scientific abstracts in the English language. In this paper we work with Russian texts. The obtained results show that the proposed approach works not only for texts that are appropriate in terms of structure and literacy, such as abstracts, but also for short texts, such as forum messages, in which many words may be misspelled and the text itself is poorly structured. Moreover, the results show that proposed approach works well not only with English texts, but also with texts in the Russian language.**

Keywords—keyphrase extraction, information retrieval, short texts, stop words

INTRODUCTION

Keyword extraction is a classic and important problem. It has a significant number of applied fields (including: search engine indexing, building semantic maps, identifying basic patterns and relationships, knowledge extraction, e.t.c.). The paper considers the problem of multi-word keyword extraction, hereinafter referred to as "key phrase extraction". The main assumption of this work is: extended lists of stop words can significantly improve the results of basic key phrase extraction algorithms.

In the previous study [1] it was shown that it is possible to improve the quality of the extracted key phrases if an extended list of stop words is used. The approach was applied to abstracts of scientific publications written in English. In the present paper we expand the approach proposed in [1] on another type of texts which are in the Russian language.

Under an extended list of stop words, we understand a list of such words that are not included in the list of standard stop words, but when adding them to the standard list the performance of the main annotation algorithm for a particular task improves (assuming the algorithm uses the lists of stop words in it's work).

The reported study was funded by RFBR according to the research project No. 16-37-00430 mol_a and partially supported by the Government of Russian Federation, Grant 074-U01.

STATE-OF-ART

In the problem of key phrase extraction two main approaches are most commonly used. The first approach considers the following steps: at the first step words are selected from the text, from which key phrases are created at the second step [2-5]. In the second approach, the candidate phrases are selected from the text at the first stage, and the key phrases are extracted from the selected candidate phrases at the second stage (typically done using ranking or classification) [4, 6-13]. We refer to [14] for a detailed and complete description of the state-of-the-art.

The presented research is carried out within the second approach, and considers only the stage of constructing the candidate phrases. It should be noted, that the quality of the extracted key phrases based on the results of the second stage (using ranking or classification) depends heavily on the quality of the phrases in the set of candidate phrases. In [6] it was mentioned that too many candidates negatively influence the ranking. Thus, it is important to construct such algorithms that will extract small sets of candidate phrases. Authors of [14], note that with the increase in the length of the document and the number of generated candidate phrases, it becomes more difficult to extract key phrases due to the bigger search space. Additionally, in study [8] it was shown that the results of the second stage can be very close to the results achieved by random selection. Thus: the more correct and less incorrect phrases are received in the set of candidate phrases, the better the result will be after ranking or classification. Therefore, as the main task, we consider the task of creating a good set of candidate phrases. The task of extracting a fixed number of phrases from a large number of candidate phrases is not considered in this paper.

The authors of [14] note that there is no systematic analysis of the major types of errors in state-of-the-art key phrase extraction systems. In [14], the authors analyze and distinguish four major types of errors, among which the most frequent is the overgeneration error. The authors write: "Overgeneration errors occur when a system correctly predicts a candidate as a keyphrase because it contains a word that appears frequently in the associated document, but at the same time erroneously outputs other candidates as keyphrases because they contain the same word" [14].

Research [15] deals with the mentioned type of error. The authors proposed an ILP (integer linear program) for keyphrase extraction. The main task is to reduce overgeneration errors by weighting sets of key phrase candidates according to their component words.

In the present research, we work with an approach, that indirectly deals with the mentioned type of error. The main idea is to use extended lists of stop words. By stop words we understand words that can not be found in the phrases that the algorithm extracts. Although in general these words are ordinary words and are not related to standard stop words. As a matter of fact, the proposed approach makes it possible to single out and tag as stop words – words that are more often found in incorrect candidates than in correct ones. Once the word is tagged as a stop word, no phrases are constructed with this word. Since the phrases with this word were mostly not correct, removing such phrases increases Precision faster than Recall decreases. Thus, we find words that lead to the overgeneration error. Deleting such words, allows to delete the set of incorrect phrases generated with this word. At the same time, a certain number of correct phrases are lost, but this loss is leveled by the gain from removing a large number of incorrect phrases. In our research, we want to show that such words can be found automatically with the help of training collections: texts, with hand-selected key phrases.

The purpose of this study is to expand and further substantiate the approach to extracting key phrases using extended lists of stop words. Also we would like to show that the proposed approach works not only for texts that are correct in terms of structure and literacy, such as abstracts of scientific publications (as in [1]), but also for very poorly structured short texts, such as Internet forum messages. The research expands the existing field of key phrase extraction. The aspects considered in this study are not reflected in other works in the field.

DATASET

A collection of 120 messages from Internet car forums was collected containing feedback and opinions on the purchase and repair of cars. Half of the texts contain positive reviews, and the other half is negative. All texts in the collection have the “gold standard”: manually assigned key phrases. The collection was marked in such a way that the key phrases reflected the main content of the text, including the names of car-care centers and salons, opinions, objects of opinion and reasons for the opinion expressed. The texts were mixed and divided into two equal collections, further denoted as the T1 collection and the T2 collection. Table 1 provides examples of texts and the assigned key phrases.

EVALUATION

A standard evaluation approach was used. Automatically extracted key phrases are compared with the key phrases assigned by an expert:

$$Precision = \frac{|(C \cap G)|}{|G|} \qquad (1)$$

TABLE I. EXAMPLES OF TEXTS AND KEY PHRASES (TRANSLATED TO ENGLISH AND ORIGINAL TEXTS)

Examples of texts	***Phrases of the "gold standard" (manually assigned phrases)***
Ordered a car - hyundai solaris, they took a prepayment of 50 thousand. rubles., we have been waiting for half a year, they do not really say anything! Disgusting work of the salon, they take not small amounts of money from customers, screw them over, and then offer to re-order the car in another salon! *заказали у них машину hyundai solaris, взяли предоплату 50тыс. руб., ждем уже пол года, ничего толком не говорят! отвратительная работа салона, берут с клиентов не малые деньги, прокручивают их, а потом предлагают перезаказать машину в другом автосалоне!*	Ordered a car; took prepayment; waiting for half a year; disgusting work of the salon; offer to re-order the car *заказали машину; взяли предоплату; ждем пол года; отвратительная работа салона; предлагают перезаказать машину*
The work took a very long time, not high-quality and expensive. Example: when adjusting the valves, I waited there for 4 hours, but the valves kept knocking as they have knocked before. *работу проводят очень долго, не качественно и дорого. пример: при регулировке клапанов я там прождал 4 часа, а клапана как стучали так и стучат.*	A very long time; not high-quality; expensive; valves kept knocking *очень долго; не качественно; дорого; клапана стучат*
Car service koreana is great. Needed an urgent replacement of the injectors. Called to sign up, was asked to come in an hour. I arrived, gave the car and the injectors - in forty minutes they returned the car (intact and safe). The prices were a pleasant surprise. *автосервис кореана- зачет. нужна была срочная замена форсунки. позвонил записаться, сказали приехать через часик. приехал, отдал машину и форсунки - через сорок минут отдали готовую(в целости и сохранности). цены приятно удивили.*	Car service koreana; great; urgent replacement of the injectors; prices were a pleasant surprise; *автосервис кореана; зачет; срочная замена форсунки; цены приятно удивили*

$$Recall = \frac{|(C \cap G)|}{|C|} \qquad (2)$$

$$Fscore = \frac{(2 \cdot Precision \cdot Recall)}{(Precision + Recall)}, \qquad (3)$$

where $|(C \cap G)|$ is the total number of correctly extracted phrases by the algorithm when processing all the texts of the collection, $|G|$ is the number of phrases automatically extracted by the algorithm from all the texts of the collection, $|C|$ is the number of all phrases from the “gold standard”.

EXPERIMENT DESCRIPTION

A. Research hypothesis

The essence of the approach can be defined as follows: there is an algorithm that extracts key phrases. Stop words are delimiters for phrases and can not be included in the key phrases. A set of words S, which needs to be included as stop words, must be determined to improve the key phrase extraction operation of the algorithm. It is assumed that the use of the same algorithm along with the selected new stop words will improve the quality of the key phrase

extraction in other similar collections. In [1] a close study was carried out for English-language texts and tested on a dataset of abstracts of scientific publications. It was shown that the proposed approach, despite its comparative simplicity, improves the quality of the basic algorithm. In this study, Russian-language texts from social media are used for processing. The two main differences between the datasets are that they are in different languages and the texts in the Russian dataset are much more poorly structured with many misspellings and slang.

It should be noted that a word can be both a stop word for some texts and be part of key phrases in others. Therefore, to justify the addition of each particular word to the list of stop words we should take into account the effect that each word has by calculating the ratio of the gain in quality (due to the fact that the phrases become more precise) to a loss in quality (since part of the correct phrases containing the added word are lost). Thus, the task is to select only such words, the addition of which to the stop words list will improve the quality of the algorithm on the whole collection by at least a specified value. The research hypothesis is that such words can be extracted from one collection (training sample) and applied to the other collection (test sample). To use this approach, one must have a collection of assigned key phrases. This may be considered as a positive aspect, since during "gold standard" creation it is possible to determine the type and "logic" of phrases that are interesting to extract (for example: should the phrases be long or short, what should they reflect (subject, opinion, location, etc.)).

B. Automatic extraction of extended list of stop words

Suppose there is an algorithm for key phrase extraction that uses a list of stop words. Using a training collection, we build an extended list of stop words. Let V be the vocabulary of the training collection. Let S be the set of words included in the extended list of stop words and let $Sbase$ be the set of words included in the standard list of stop words of the selected language. Let $Fscore(S)$ evaluate the quality of the algorithm, which uses a list of stop words S. The extended list of stop words is extracted by the following algorithm:

$S=\varnothing$

for $\forall v \in V$*:*

{if Fscore(v∪Sbase)-Fscore(Sbase)>p=>v∪S}

$S=S\cup Sbase$

The obtained extended list of stop words S is used by the main algorithm for processing the test collection.

C. Algorithm for extracting phrases

Two approaches are considered. The first approach consists in extracting phrases that are sequences of words of given parts of speech, where the sequences are as long as possible. At the same time, words of distinct parts of speech, punctuation and stop words are used as separators. The approach showed very good results as described in [1], [8].

The second approach is based on key phrase extraction with the help of linguistic patterns, which are sequences of parts of speech that must correspond to words in the phrase. To build such patterns the "gold standard" of the training collection is used. We select the most frequent patterns matching the phrases of the "gold standard". These patterns are used to extract key phrases from both training and test collections.

It must be noted that key phrases extracted using patterns, in contrast to phrases extracted as maximum word sequences of certain parts of speech, have higher Recall and lower Precision. Which can be explained by the following: when using patterns, more phrases are extracted than when using sequences. For example, if the patterns given are: "adjective + noun" and "adjective + adjective + noun," then, when using the patterns approach two phrases will be extracted, where one phrase will be part of the other. But only one phrase "adjective + adjective + noun" will be extracted from the text using the longest sequence approach.

Selection of patterns.

Sequences of parts of speech (patterns) were identified using the gold standard of the training collection. Only patterns that occur at least k times were selected.

For example, let $T1$ and $T2$ be the training and test collections accordingly. Patterns are extracted from gold standard of $T1$. They are used to extract potential phrases for texts from collections: 1) $T1$, when building an extended list of stop words; 2) $T2$, when evaluating the quality of the algorithm with the obtained lists of stop words.

Selection of parts of speech.

For the case when phrases are extracted from the text as maximum continuous sequences of words of certain parts of speech, only those parts of speech that are found in the patterns obtained in the "Selection of patterns" section were chosen. In all experiments the following parts of speech were used: nouns, adjectives and verbs.

Texts preprocessing.

All texts in the training and test collections, as well as their gold standards, were pre-processed: lemmatized and part-of-speech tagged using MyStem [16] with the following changes. Since the texts often discuss the maintenance of cars commonly used abbreviations, as "то" and "авто", which stand for "vehicle inspection" and "car" in the Russian language, are tagged as a nouns during POS-tagging. Also, separately with a unique tag, the "не" particle (translated as "not") is allocated.

EXPERIMENT RESULTS AND DISCUSSION

The results are presented in Table 2. The table show results for the cases, when key phrases were extracted as continuous sequences of words of certain parts of speech and using patterns.

The results are given for the cases:

- where the $T1$ collection was used to build an extended list of stop words, and the $T2$ collection was used to test the extracted lists;
- for the opposite cases, where the $T2$ collection was used to extract the extended list of stop words, the $T1$ collection was used for testing.

The following denotations are used in Table 2. The denotation "Patterns" is used for the results of the algorithm that extracts phrases based on the patterns (for more details, see Section V). The denotation of "Sequences of words" is used for the results of the algorithm that extracts phrases based on sequences (more in detail, see Section V).The denotation $T1 => T2$ is used for the case when the extended list of stop words was extracted from the $T1$ collection, and the quality evaluation of the algorithm was performed on the $T2$ collection. The denotation $T2 => T1$ is accepted for the opposite case.

Patterns used in the $T1 => T2$ case: s, a_s, v_s, a, s_s, v, s_v, s_p_s.

Patterns used in the $T2 => T1$ case: s, a, a_s, v_s, s_s, v.

Where $s -$ is a noun, $a -$ an adjective, $v -$ a verb, $p -$ a preposition.

In Table 2 it is shown that the use of extended lists of stop words allows to improve the quality of the results on the test collection. If the threshold value p is decreased, more words are added to the extended list of stop words and this leads to an improvement in the quality of the results.

Loss in the quality of extraction of key phrases when using extended lists of stop words is not observed in any of the cases. This allows to assume that the same extended lists of stop words can be applied to different collections of the same type of documents. The latter indicates some universality of words that are more often not found than are found in key phrases within a set of similar documents.

TABLE II. EVALUATION OF THE QUALITY OF THE EXTRACTED KEY PHRASES FOR THE TEST AND TRAIN COLLECTIONS: STANDARD AND EXTENDED LISTS OF STOP WORDS ARE USED

Stop words	**Standard list of stop words**	**Extended list of stop words**		
p-value		***0.0003***	***0.0001***	***0.00005***
Patterns				
T1=>T2	0.18	0.20	0.21	0.21
T2=>T1	0.20	0.19	0.20	0.21
Sequences of words				
T1=>T2	0.24	0.24	0.24	0.24
T2=>T1	0.24	0.25	0.27	0.27

CONCLUSION

In this research we have set several goals. The first goal was to expand and further justify the approach to extracting key phrases using extended lists of stop words. The task was to show that the constructed extended lists of stop words for one collection can be used for similar collections and can improve the quality of the extracted key phrases. By similar collections we mean texts that are similar to the texts in the training sample or from similar sources.

The second goal was to show that the proposed approach works not only for texts that are sufficiently correct in terms of structure and literacy, such as abstracts of scientific publications (as in [1]), but also for very poorly structured short texts, such as Internet forum messages, in which, among other things, many words are misspelled or slang. The task was to test experimentally whether it is possible to improve the operation of the basic algorithm with the help of an extended list of stop words for Russian-language messages from Internet forums.

In the experiments, positive results were obtained. This allows to confirm the possibility of improving the work of algorithms that extract key phrases using extended lists of stop words obtained from the training collection.

REFERENCES

[1] S. Popova, L. Kovriguina, D. Muromtsev and I. Khodyrev, "Stop-words in Keyphrase Extraction Problem," in Proc. of 14th Conference of Open Innovations Association FRUCT. Helsinki, Finland, 2013, pp. 113–121.

[2] R. Mihalcea, P. Tarau TextRank: Bringing order into texts // Proc. of the Conference on Empirical Methods in Natural Language Processing, 2004, pp. 404–411.

[3] W. Xiaojun and J. Xiao "Single document keyphrase extraction using neighborhood knowledge", Proceedings of the 23rd AAAI Conference on Artificial Intelligence, 2008, pp. 855–860.

[4] T. Zesch, I. Gurevych, "Approximate Matching for Evaluating Keyphrase Extraction", International Conference RANLP 2009, Borovets, Bulgaria, 2009, pp. 484–489.

[5] K. S. Hasan and V. Ng, "Conundrums in unsupervised keyphrase extraction: Making sense of the state-of-the-art.", Proc. Of the 23rd International Conference on Computational Linguistics: Posters, pp. 365–373, 2010.

[6] W. You, D. Fontaine and J.-P. Barhes, "An automatic keyphrase extraction system for scientific documents," In: Knowl Inf Syst 34, 2013, pp. 691-724.

[7] S. R. El-Beltagy and A. Rafea "KP-Miner: A keyphrase extraction system for english and arabic documents," in: Information Systems, 34, 2009, pp. 132-144.

[8] S. Popova, I. Khodyrev "Ranking in keyphrase extraction problem: is it useful to use statistics of words occurrences?", Proceedings of the Institute for System Programming of the RAS, 2014, №4

[9] S. N. Kim, O. Medelyan and M. Yen, "Automatic keyphrase extraction from scientific articles," Language Resources and Evaluation, Springer Kan & Timothy Baldwin, 2012

[10] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin and C. G. Nevill-Manning "Domain-specific keyphrase extraction," in: Proc. of IJCAI, 1999, pp. 688–673.

[11] P. Turney "Learning to Extract Keyphrases from Text," in: NRC/ERB-1057,1999, pp. 17– 43.

[12] A. Hulth "Improved automatic keyword extraction given more linguistic knowledge," in: Conference on Empirical Methods in Natural Language Processing, pp. 216–223, 2003

[13] S. Popova, G. Skitalinskaya, I. Khodyrev, "Estimating Keyphrases Popularity in Sampling Collections", Ciuciu I. et al. (eds) On the Move to Meaningful Internet Systems: OTM 2015 Workshops. Lecture Notes in Computer Science, vol 9416. Springer, Cham.

[14] K. S. Hasan and V. Ng Automatic Keyphrase Extraction: A Survey of the State of the Art. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp 1262–1273, Baltimore, Maryland, USA, 2014

[15] F. Boudin, "Reducing Over-generation Errors for Automatic Keyphrase Extraction using Integer Linear Programming", Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction, Beijing, China, 2015. p. 18.

[16] MyStem, https://tech.yandex.ru/mystem/