A Generalized Approach to Keyphrase Extraction using Extended Lists of Stop Words

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

Gabriella Skitalinskaya
Institute of Technology Tallaght
Dublin, Ireland
gabriellasky@icloud.com

***Abstract*—The article presents a generalized approach for keyphrase extraction based on the construction of extended lists of stop words. A description of the approach is given, as well as generalization of observations made within the framework of its development.**

I. Introduction and approach description

Keyphrase extraction is an important problem of natural language processing, which suffers from poor performance relative to many other core natural language processing problems. The most complete overview of the current state-of-the-art is represented in [1]. "SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles" are presented in [4]. To get acquainted with the state-of-the-art, one should also get acquainted with such resources as: SemEval 2017 Task 10: Extracting Keyphrases and Relations from Scientific Publications [2], the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction [3].

The approach to extracting keyphrases developed in our works [4], [5], [6], [7] complements and expands the research in the field and represents a separate direction. It is based on the use of extended lists of stop words. By stop words we understand words that can not be found in keyphrases and at the same time are separators between the phrases. Thus, the proposed approach extracts candidate phrases as continuous word sequences of maximum length of given parts of speech. Delimiters between phrases in this case are punctuation marks, words of other parts of speech and stop words. The extended list of stop words contains in addition to the standard stop words for the language, other words, which are common words of the language. In this paper, we justify this algorithm and generalize the observations obtained within the framework of this approach.

II. Essence of the approach

The most popular approach in the area of keyphrase extraction consists of two stages: the construction of candidate phrases and their further classification or ranking. Note that the task of extracting keyphrases is complicated by the fact that the frequency values of words that are in keyphrases and words which are not, can practically not have differences [4]. Another problem is that the word can be simultaneously a stop word for some phrases and a keyword for other phrases. Moreover, the same phrase can be keyphrases for one text, and not for another text. This complicates the classification / ranking process. We also note the following [9]: a large number of generated candidate phrases extremely negatively affects the quality of selection of keyphrases at the ranking stage. The paper also shows that the use of information about the proximity of the phrase to the beginning of the document works well in the case of scientific publications and does not work for literary texts. This is an interesting observation, since one of the main criteria in assessing the weight of a phrase is the position of the phrase in relation to the beginning of the text.

The approach developed by us allows us to initially improve the quality of the extracted candidate phrases, which could additionally be classified or ranked in the next stage. The essence of the proposed approach is to remove such phrases from candidate phrases the words of which are more often not found in keyphrases than found. We propose to add such words to the extended lists of stop words and use these lists during the extraction of phrases. It is important to consider how the addition of each particular word to the list of stop words is justified: the ratio of the gain in quality (due to the fact that the phrases become more precise) to loss in quality (since part of the correct phrases containing the added word is lost). To obtain such words, a training collection is used (for a detailed description, see [5]). The training collection is a set of texts for which keyphrases are already known and marked by the experts manually (hereinafter "the gold standard of the collection.") In [5] extended lists of stop words allowed us to improve the quality of the extracted phrases for English scientific abstracts. Similar results, but for texts in the Russian language from Internet car forums, were achieved in [6], [7].

III. Algorithm for constructing phrases

Along with the algorithm which extracts phrases from a text as the longest sequences of words from given parts of speech (delimiters: punctuation marks, words of other parts of speech and stop words) we apply the extended lists of stop words. This approach works for documents of different types: scientific publications and Internet forum messages [4], [5] and [6], [7] and for documents in different languages (for example, Russian and English languages). To determine the list of parts of speech to be used during extraction, the training collection and its gold standard are used. On the basis of the training collection the most frequent linguistic patterns are extracted to which the keyphrases of the gold standard correspond. By linguistic pattern we understand a certain predetermined sequence of parts of speech. In the next step, only those patterns from the resulting set of patterns are left, which: 1) most often correspond to the phrases of the gold standard of the training collection, 2) and for these patterns a high ratio tp/fp

is obtained, where tp (true-positive) is the number of phrases correctly extracted using this pattern, and fp (false-positive) is the number of phrases retrieved using the patterns, which are not keyphrases. Further, patterns, for which tp/fp is high, will be called maximally effective patterns. Parts of speech that are found in the frequent maximally effective patterns of a training collection's gold standard are used to construct the mentioned sequences.

The use of sequences in our opinion is more appropriate than using frequent gold standard's patterns due to the following:

1) From the point of view of the F1-score quality assessment, the use of sequences experimentally yields better results than the use of patterns. The computational approach is simpler than using patterns. The higher F1-score is achieved due to greater Precision of the selected phrases in comparison with the case of using a pattern-based approach. Although phrases extracted using patterns have a better Recall, in the case of sequences, this difference is compensated for by a higher Precision value.
2) The pattern-based approach requires careful selection of patterns suitable for use. The most frequent patterns used in the gold collection standard require additional filtering and are not always appropriate. The problem is that such patterns can extract not only keyphrases, but also a large number of phrases that are not keyphrases. This leads to severe loss in Precision.
3) In general, the effectiveness of a sequence-based approach is a consequence of a patterns-based approach. Really if you select frequent the most effective patterns from a training collection's gold standard, then such patterns will be a combination of the parts of speech that are best suited for constructing phrases. For example, in the case of annotations to scientific publications, these are nouns and adjectives. The latter provides a good work of the algorithm, extracting from the annotations of the phrase as the maximum sequences of nouns and adjectives. In the case of messages from Internet car forums, the main parts of the speech are: nouns, adjectives and verbs. These same parts of speech are frequent, most effective templates for the collection of messages from the forums.

We add one more remark. When extracting phrases as the longest sequences, the question arises whether to include / exclude single-word phrases. Earlier we showed the effectiveness of deleting single-word phrases when annotating scientific publications [4], [5]. But for other types of collections (for example, for social media texts) deleting single-word phrases can have the opposite effect. The reason is similar to that described in clause 3 and is that for abstracts one-word patterns are not effective, but for messages from forums they are. Also the gold standard collection of forum posts contains more single-word phrases, the texts themselves have a significantly larger number of punctuation marks per unit text.

IV. Conclusion

In this paper we summarized and presented our research in the field of keyphrase extraction as a single approach. The comparative simplicity of the proposed approach allows to receive very good results in field (from the point of view of the F1-score, e.g. F1=0.445 for INSPEC collection [5]). The essence of the approach is to combine an algorithm that extracts phrases as continuous sequences of words of given parts of speech with the use extended lists of stop words. A method is defined for obtaining the required list of parts of speech and the list of extended stop words based on the training collection.

References

[1] Hasan, K.S., Ng, V.: "Automatic Keyphrase Extraction: A Survey of the State of the Art", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* 2014, pp. 12.

[2] Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: "Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications", *CoRR abs/1704.02853* 2017

[3] ACL: "The 53rd Annual Meeting of the Association for Computational Linguistics.", *In: Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction* 2015

[4] Popova, S., Khodyrev, I.: "Ranking in keyphrase extraction problem: is it suitable to use statistics of words occurrences? ", *Proceedings of the Institute for System Programming* 26, 2014 , pp. 123-136.

[5] Popova, S., Kovriguina, L., Mouromtsev, D., Khodyrev, I.: "Stop-words in keyphrase extraction problem", *In: Conference of Open Innovation Association, FRUCT*, 2013 , pp. 113-121.

[6] Popova, S., Skitalinskaya G.: "Extended List of Stop Words: Does It. Work for Keyphrase Extraction from Short Texts?", *In: Intern. Conf. CSIT-2017, IEEE*, 2017 , 4 pp. [in press]

[7] Popova, S., Skitalinskaya G.: "Keyphrase Extraction Using Extended List of Stop Words with Automated Updating of Stop Words List", *In: Advances in Intelligent Systems and Computing, Springer*, 2017 , 13 pp. [in press]

[8] Hulth, A.: "Improved Automatic Keyword Extraction Given More Linguistic Knowledge", *Language* 2003, pp. 216223.

[9] You, W., Fontaine, D., Barths, J.P.: "An automatic keyphrase extraction system for scientific documents ", *Knowledge and Information Systems* 34, 2013, pp. 691-724.