THE STUDY OF FACTORS RELADED WITH SINGLE-DOCUMENT KEYWORD EXTRACTION

Popova Svetlana

***Abstract**: In this paper we consider the problem of keyword extraction for the document annotation. We study different statistic-based measures for such extraction and compare the results with the annotations made by skilled experts. The experiments were completed with the Inspec data set. The quality of annotation was evaluated by means of R-Precision score.*

***Keywords**: keywords extraction, single document annotation, natural language processing.*

***ACM Classification Keywords**: I.2.7. Natural Language Processing*

Introduction

The aim of keyword extraction is to find a small number of terms representing the document content "in a few words". Keyword extraction systems may be useful in automatic texts classification, clustering, document indexing, novelty detection, question-answering, text representation etc. In this paper we focus on the problem of keyword extraction for the purpose of document annotation. Our approach consists of two steps. The first step includes token selection and their grouping into multi-words candidates. At the next step we rank all candidates and then top-ranking candidates are selected as keywords. We tested efficiency of various statistics for these procedures. It was experimentally shown that it is more preferable to remove candidates having one-word-length than to use any statistics from a given list. We used R-Precision for final evaluation.

State-of-the-art

The majority of keyword extraction approaches are statistical-based and graph-based. P. Turney [Turney, 2000] proposed a supervised system that combined genetic algorithm with the parameterized keywords extraction system Extractor. E. Frank [Frank, 1999] used supervised system based on a Naive Bayes classifier. A. Hulth [Hulth, 2003] combined learning process and linguistic information to deal with keyword extraction problem. In her work it was shown that POS tagging improves all results independent of the applied term selection approach. Unsupervised graph-based approach was proposed by [Mihalcea, 2004]. In this approach text entities were presented as vertexes linked with each other. The scores of the graph vertexes were computing using formula based on PageRank. X. Wan [Wan, 2008] explored neighborhood knowledge to score co-occur statistics to TextRank. T.Zesch [Zesch, 2009] proposed the generalized framework for comprehensive analysis of keywords extraction that included different combinations of candidate construction and candidate ranking.

The analysis of the state-of-the-art shows that the problem of keyword extraction is far from its solution. The obtained results are not high and so the problem should be elaborated more carefully. Another difficult question is evaluation of results quality. New measures for quality evaluation were proposed in [Zesch, 2009; Su Nam Kim, 2010]. But by the moment there is no experience in application of these measures. In this paper we use the quality measures proposed in [Zesch, 2009], and compare our results with those described in [Zesch, 2009].

Dataset

This research is based on experiments with Inspec, one of the main datasets mentioned in the state of the art review. It contains 2000 abstracts from Computer science and Information Technology and consists of three subsets: trial set (1000 abstracts), validation set (500 abstracts) and test set (500 abstracts). As in all previous works we based our research on the test subset. Every document in Inspect has a gold standard, which was created by experts. The gold standard includes two sets of keywords: "contr" set and "uncontr" set. We used "uncontr" set. More detailed information about the collection could be found in the paper [Hulth, 2003].

Evaluations

Previously, the method based on *F–score* [Manning, 2009] was used to evaluate results in the keywords extraction problem. But there was a problem to understand how many keywords should be extracted. Recently it has been proposed to use *R-Precision* (R-p) [Zesch, 2009]. R-p is Precision when the number of extracted keywords is equal to the number of keywords in the gold standard. R-p allows us to consider the problem of keywords extraction as the problem of ranking, when candidates to keywords should be ranked in order to detect the most important among them. Another question is how to understand that extracted keyword k is correct. T. Zesch [Zesch, 2009] considered k correct if it overlapped the keyword g from the gold standard or if k and g were morphological variants of each other. In our research we use R-p to evaluate results in two cases: 1) exact: k consider correct if it equals to g in each word, 2) include: k is correct if it overlaps g .

Algorithm

The pre-processing step included: stop words removing, splitting text into sentences and part of speech tagging (Stanford POS tagging tool). All words except for nouns and adjectives were removed from texts. For each text three statistics were exploited separately to select tokens. These statistics were: *tf-idf* [Manning, 2009], *within document term frequency* and *Transition Point* which has been successfully used for the term selection in the clustering problem [Pinto, 2006]. Tf-idf requires information from all documents in a collection. Other two statistics work with single documents. Transition Point technique is based on the idea that mid-frequency terms are semantically close to the text content. Basically, formula of this technique that calculates TP_d for the document d is:

$$TP_d = \frac{\sqrt{8 * I_1 + 1} - 1}{2},$$

I_1 is the number of words with the frequency equal to 1 in *d*. All words with-within document term frequency *wdf* are selected from d if:

$$U_1 \le wdf \le U_2,\ U_1 = (1 - c) * TP_d,\ U_1 = (1 + c) * TP_d,\ 0 \le c < 1.$$

We grouped selected tokens that followed each other in the original document to construct multi-words candidates. Most of keywords contain 1 to 4 words and we reconstructed candidates not longer than four words [Zesch, 2009]. The next step was ranking process. The final score for the candidate was calculated as average value of all the contained tokens. We explored three statistics to score token's value: 1) *tf-idf*, 2) within document frequency, 3) transition point (that was calculated as: $tp_{token} = | TP_d - wdf_{token} |$). Additionally we used average mutual information (*MI*) for ranking candidates [Manning, 2009] calculated between all words pairs in a candidate. If a candidate included only one word its score was equal to zero.

Experiment

We compared all combinations of the token selection methods (tf-idf, within document frequency (*wdf)*, transition point (*tp*)) and the candidates ranking criteria (tf-idf, within document frequency (*wdf*), transition point (*tp*), MI). Experiments showed that the best results were obtained if all tokens were selected independently of the token selection strategy. Additionally in experiments we tried to remove words that occur in the text only once, but it reduced quality. Results of experiments with four candidates ranking methods present in Table 1 a) assuming that all terms were selected on the term selection stage. MI showed the best results because scores of all one-word-long candidates were calculated as zero. The majority of these candidates were not included in the keywords list because the values of these candidates were minimal. In the next experiment we changed ranking step for every criterion: firstly all one-word-long candidates were removed and then all other candidates were ranked. Table 1 b) presents results for this experiment and it shows that all ranking strategies are equal, except MI having lower result. Comparing Table 1 a) and b) allows us to draw an interesting conclusion: deletion of one-word-long candidates improves results and has more influence on quality than any of the performed ranking criteria. It could mean that content words rather follow each other in the text frequently than appear alone. Evaluation with R-p (include) 0.37 outperforms all results for Inspect dataset present in the work [Zesch, 2009].

Table 1. Experiment Results a) all candidates, b) without one-word-long candidates

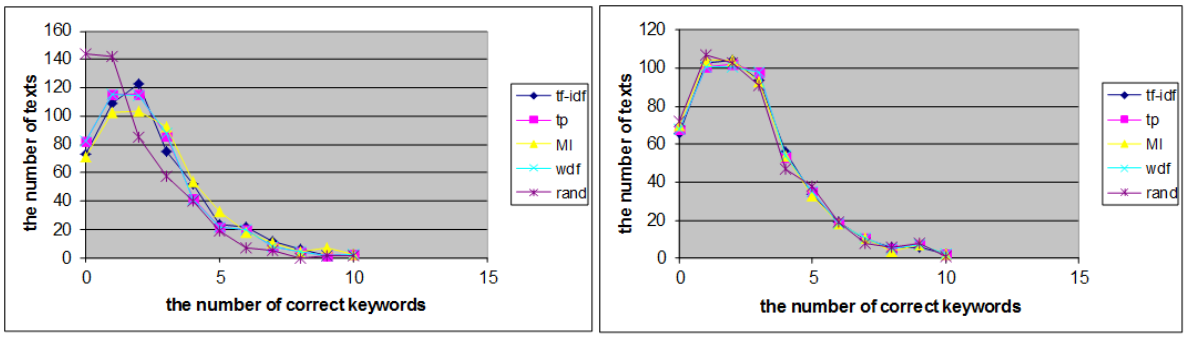
all candidates	R-p (exact)	R-p (include)
tf-idf	0.25	0.32
transition point	0.23	0.30
MI	0.28	**0.36**
wdf	0.23	0.30

a)

except one-word-long candidates	R-p (exact)	R-p (include)
tf-idf	0.28	**0.37**
Transition point	0.28	**0.37**
MI	0.28	0.36
wdf	0.28	**0.37**

b)

In the next step we explored why all ranking methods gave the same results in Table 1 b). We selected keywords randomly (*rand*) from candidates assuming that all one-word-long candidates had been removed. The result was positive: R-p (exact) 0.28, R-p (include) 0.37. It happened because in the major part of cases the number of constructed candidates for the text was almost the same as the number of keywords in the gold standard. However if we do not remove one-word-long candidates, random selection of keywords from candidates will reduce quality: R-p (exact) 0.17, R-p (include) 0.22. Diagram 1 a) and b) shows dependence between the number



a) *b)*

Fig. 1. a) all candidates, b) without one-word-long candidates

of documents and the number of correct extracted keywords per document for different ranking strategies (case: exact). Diagram 1 shows that removing one-word-long candidates allows increasing the number of correct extracted keywords per document and all ranking strategies do it in the same way (include random selection). It proved that candidates to keywords with one-word length bring noise in ranking step.

Conclusion

In this paper we studied the influence of various statistic methods on keyword extraction problem. We propose the approach related with this problem, which includes token selection and candidate ranking steps. The result is that the difference between performed measures in proposed approach is poor. If we remove all words from text except for nouns and adjectives, group the remained words into multi-words candidates, and then remove all one-word-long candidates, no ranking will be needed. It means that additional information and other measures or methods are needed to improve results. Removing from the text words that only once occur in it reduces quality. It proved that some of these words were included in keyword list. We detected that deletion of one-word-long candidates during ranking step improves results. It means that content words rather follow each other in the text frequently than appear alone. In the future we suppose to study graph-based and knowledge-based measures.

Bibliography

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

[Hulth, 2003] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In: Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 216–223, 2003.

[Manning, 2009] C. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University , 2009.

[Mihalcea, 2004] R. Mihalcea, P. Tarau. TextRank: Bringing order into texts. In: Proc. of the Conference on Empirical Methods in Natural LanguageProcessing, pages 404–411, 2004.

[Pinto, 2006] D. Pinto, H. Jim´enez-Salazar, P. Rosso. Clustering Abstracts of Scientific Texts Using the Transition Point Technique. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 536–546. Springer, Heidelberg, 2006.

[Su Nam Kim, 2010] Su Nam Kim, T. Baldwin, Min-Yen Kan. Evaluating N-gram based Evaluation Metrics for Automatic Keyphrase Extraction. In: Proc. of the 23rd International Conference on Computational Linguistics, pages 572–580, 2010

[Turney, 1999] P. Turney. Learning to Extract Keyphrases from Text. Published as NRC/ERB-1057, 43 pages, 1999.

[Wan, 2008] X. Wan, J. Xiao. Single document keyphrase extraction using neighborhood knowledge. In: Proc. of the 23rd AAAI Conference on Artificial Intelligence, pages 855–860, 2008.

[Zesch, 2009] T. Zesch, I. Gurevych. Approximate Matching for Evaluating Keyphrase Extraction. In: Proc. International Conference RANLP 2009 - Borovets, Bulgaria, pages 484–489, 2009.

Authors' Information

Svetlana Popova *– Saint-Petersburg State University, the Department of Programming Technology, PhD program student; e-mail: spbu@bk.ru*

Major Fields of Scientific Research: documents clustering, scientific documents processing