ENGINEERING IN ADVANCED
RESEARCHSCIENCE AND
TECHNOLOGY

ISSN 2352-8648
Vol.01, Issue.03
April-2019
Pages: -169-177

AUTOMATIC KEY PHRASE EXTRACTION FOR MULTI DOCUMENT

A. BRAHMA REDDY[1], DR. VAKA MURALI MOHAN[2], DR. KANAKA DURGA RETURI[3]

[1] PhD scholar, Chowdary Charan Singh University, Meerut, UP.

[2] Professor, Department of Computer Science and Engineering, Malla reddy College of engineering for women, Misammaguda, Medchal, Hyderabad, TS.

[3] Professor, Department of Computer Science and Engineering, Malla reddy College of engineering for women, Misammaguda, Medchal, Hyderabad, TS.

Abstract: Key phrase extraction only consider the connections between words in a document, ignoring the impact of the sentence. Motivated by the fact that a word must be important if it appears in many important sentences, we propose to take full advantage of the reinforcement between words and sentences by melting three kinds of relationships between them. Moreover, a document is grouped with many topics. The extracted key phrases should be synthetic in the sense that they should deal with all the main topics in a document. Inspired by this, we take topic model into consider. Experimental results show that our approach performs better than state-of-the-art key phrase extraction method on two datasets under three evaluation metrics.

Keywords: Key phrase extraction; graph-based; cluster;

1. INTRODUCTION: Keywords give a high level summarization of a document, which is vital for many areas of natural language processing, including document categorization, clustering and classification1. Before the emergence of the technology of automatic key phrase extraction, the task is usually conducted by human which is very time-consuming. Moreover, the scale of Information is becoming larger owing to the development of the Internet and it is inefficient for professional human indexers to note documents with key phrases manually. How to automatically extract the exact key phrase of a given document becomes an important research problem and many approaches have generally appeared. The task of key phrase extraction usually conducts in two steps : (1) extracting a bunch of words serving as candidate key phrases and (2) determining the correct key phrases using unsupervised or supervised approaches. In the unsupervised approach, graph-based ranking methods perform the best3. These methods construct a word graph based on word co-occurrences within the document firstly and then ranking the words according to their scores. As a result, the top ranked words are the key words we want. However, this method just maintains a single score of a word without considering the impact of sentence which is composed of words.

Copyright @ 2019**ijearst. All rights reserved.**

Motivated by the work ofWan4, we propose to extend the word graph to three graphs, namely word to word graph, sentence to sentence graph and sentence to word graph. However, Wans method has a disadvantage: the same as Text Rank

2 TERM CLUSTERING:

In this paper, we concentrate on the task of key phrase extraction other than document summarization so we just need to apply clustering techniques on the word graph. That does not mean we do not take consideration of the effect of sentence to words because only when the iteration converges and the ranking of the words are obtained, the clustering method will be applied to the graph. After clustering, we can get several clusters and each cluster contains a bunch of words which are similar to a certain topic. Then we can select the words near the centroid of each clusters key phrases according to the importance we just obtained in above sections. Here we use the widely used clustering algorithms K-means13 to cluster the candidate key phrase based on the word graph we build. The number of clusters is decided by the length of the document and we will explore the preference

3. Experiments

Datasets We carry on our experiment on two standard datasets to evaluate the performance of our method. One dataset is built by Hulth200314 and this dataset contains 1460 abstracts of research articles. Each abstract has two kinds of manually labeled key phrases, one controls the limit the other does not. Another dataset is created by Luis Marujo who annotates the articles collected from the Web with a ranked list of key phrases. The corpus contains 900 articles and each article has a list of key phrases. We call this dataset 500N in this paper

3. 1: Comparing results on Hulth2003 datasets.

System	Precision	Recall	F-measure
TF - IDF	32.8	33.0	31.4
Text Rank	39.7	40.1	37.3
Our Method	43.0	40.2	39.6

3. 2: Comparing results on 500N datasets

System	Precision	Recall	F-measure
TF - IDF	43.8	43.8	41.1
Text Rank	48.6	50.0	47.7
Our Method	48.7	49.8	47.8

from the results, our method outperforms the baseline approaches significantly. Seen from table 2, although the recall of Text Rank is better than the results of ours but our F-measure and Precision are better than it. Because we fail to get the number of key phrase extracted from each article by Text Rank, 4.Typical workflow:

To use this feature, you submit data for analysis and handle the API output in your application. Analysis is performed as-is, with no additional customization to the model used on your data.

1.Create an Azure Language resource, which grants you access to the features offered by Azure Cognitive Service for Language. It will generate a password (called a key) and an endpoint URL that you'll use to authenticate API requests.

2.Create a request using either the REST API or the client library for C#, Java, JavaScript, and Python. You can also send asynchronous calls with a batch request to combine API requests for multiple features into a single call.

3.Send the request containing your data as raw unstructured text. Your key and endpoint will be used for authentication.

4.Stream or store the response locally. The result will be a collection of recognized entities in your text, with URLs to Wikipedia as an online knowledge base.

4. Conclusions and Future work:

In this paper we propose a graph-based method incorporating with clustering algorithm for key phrase extraction . Experiments show that our method outperforms other baseline methods on two datasets. Compared with the old work conducted by Wan, we apply the clustering algorithm on the word graph and obtain an improved result which can cover the main topics the former fails to do.

5.Next some other works can be conducted on this task:

1. The clustering method can be modified due to the appearance of many clustering methods which outperforms

K-Means such as Affinity Propagation, hierarchical clustering method.

2. The method only takes consider in a single document next we can make full use of corpus which has a bunch of documents similar to the specific document.

6. References

1. C.D. Manning and H. Schutze. Foundations of statistical natural language processing. MIT Press. 2000)

2. Kazi Saidul Hasan and Vincent Ng.. Automatic Keyphrase Extraction: A survey of the State of the Art. (2015)

3. Mihalcea and P. Tarau. 2004. TextRank: Bringing order into texts. In Proceedings of EMNLP(2004) 255Yan Ying et al. / Procedia Computer Science 107 (2017) 248 – 255

4. Xiaojun Wan and Jianguo Xiao. CollabRank: Towards a collaborative approach to single-document keyphrase extraction. In Proceedings of the 22nd International Conference on Computational Linguistics, pages 969–976.(2008a)

5. Liu, Z., Li, P., Zheng, Y., Sun, M.: Clustering to find exemplar terms for keyphrase extraction.

In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing 257–266 (2009)

6. Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. Domain-specific keyphrase extraction. In Proceedings of 16th International Joint Conference on Artificial Intelligence, pages 668–673(1999)

7. Gerard Salton and Christopher Buckley. Termweighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523)(1988)

8. Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. Automatic keyphrase extraction via topic decomposition. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 366–376.(2010)

9. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine.Computer Networks, 30(1–7):107–117(1998)

10. Hongyuan Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. InProceedingsof 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 113–120(2002)

11. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research (JMLR),(2011)

12. R. Collobert. Deep Learning for Efficient Discriminative Parsing, in International Conference on Artificial Intelligence and Statistics(AISTATS), (2011)

13. Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Applied Statistics, 28(1):100-108.(1979)

14. Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of EMNLP, pages 216–223(2003)

15. Xiaojun Wan,Jianguo Xiao. Single Document Key phrase Extraction Using Neighborhood Knowledge. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence(2008)

16. Luis Marujo, Anatole Gershman, Jaime Carbonell, Robert Freder king, João P. Neto, Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12), Istanbul, Turkey, May 2012

17. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research (JMLR), (2011)

18. Brendan J J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. Science.(2007)

19. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International
Journal on Artificial Intelligence Tools 157–169 (2004)

20. Grineva, M.Grinev, M., Lizorkin, D.: Extracting key terms from noisy and multi theme documents. In: Proceedings of the 18th

International Conference on World Wide Web, pp. 661–670. ACM (2009)

21. Liu, Z., Chen, X., Zheng, Y., Sun, M.: Automatic key phrase extraction by bridging vocabulary gap. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning pp. 135–144 (2011)

22. Collo bert R, Weston J, Bottou, et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research,

Copyright @ 2019**ijearst. All rights reserved.**