# *Graph-Based Technique for Extracting Keyphrases In a Single-Document (GTEK)*

*Mahmoud R. Alfarra*
Computer Science and Information Technology Department
University College of Science and Technology
Khan Younis, Palestine
m.farra@cst.ps

*Abdalfattah M. Alfarra*
Computer Science and Information Technology Department
University College of Science and Technology
Khan Younis, Palestine
ab.alfarra@cst.ps

***Abstract*— In this paper, a novel Graph-based Technique for Extracting Keyphrases in a single document (GTEK) is introduced to be used in extractive summarization of text. GTEK is based on the graph-based representation of text, which depends on terms and phrase numeration in sentences rather than some structural document features. GTEK considers the impact of the sentence on the phrases in a document, motivated by the fact that a phrase may be important if it appears in the most important sentences in the document. The Graph-based Growing Self-Organizing Map (G-GSOM) is used to group the sentences into graph-based clusters. TextRank algorithm is applied on graphs of clusters under the assumption that the top-ranked nodes should represent the most important sentences, where the most frequent phrases in these sentences are selected as document keyphrases. Experimental results show that our innovative technique extracts the most keyphrases of two datasets.**

***Keywords*— *Keyphrase extraction; graph-based; clustering; single-document summarization.***

## I. INTRODUCTION

Keyphrases or keywords are a set of important terms in a text that give a high-level description of its content [1]. The Keyphrase extraction process is improves the content of the document with the keywords and/or key phrases explicitly mentioned in the text. The existence of the words in the document is being verified to determine the most descriptive. In general, extracting the access keyphrase does not require a third-party pre-existing to infer keywords for example glossary or related corpus.

Extracting keyphrases from a text in single or multi-documents is a significant issue that widely known in field of Information Retrieval (IR), Text Mining (TM), and Natural Language Processing (NLP) [2]. It is an elementary step for several tasks for example document clustering, document summarization and information retrieval. However, many documents do not include assigned keyphrases.

Due to the fact that assigning keyphrases manually to many documents is a costly, time consuming and boring task, especially with the appearance of big-data, constructing an effective model for keyphrase extraction becomes even more urgent and demanding at the same time [3]. Therefore, automatic keyword extraction has attracted the interest of researchers over the last years.

Keyword assignment methods can be split up into two categories as introduced in [5]:

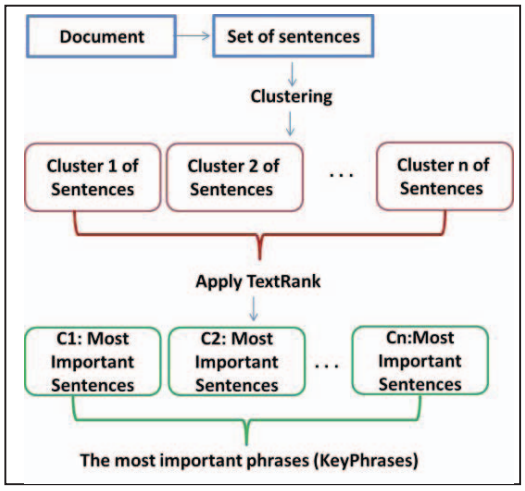
(1) Keyword assignment and;
(2) Keyword extraction.

The task of keyphrase extraction is usually carried out in two steps [4]. The first, extracting a group of words to serve as candidate keyphrases. The Second, determining the correct keyphrases using unsupervised or supervised methods.

Graph-based text representation is known as one of the best unsupervised methods to extract keyphrases. The graph is a data structure which enables inquiry of the connections between terms and structural information very effectively. Graph model is used in GTEK to represent the documents preface to extract the keyphrases taking into account some structural document features, consequently making it more accurate than the traditional vector-space model.

Advanced approaches for keyphrase extraction have a shared disadvantage, that they do not ensure the extracted keyphrases will cover all main sub-topics. Motivated by this point, a clustering-based approach is proposed in this paper to group similar sentences in groups as sub-topics of a document then extract the main important keyphrases in each cluster. This ensures that keyphrases are selected from every main sub-topic cluster. Figure 1 illustrates GTEK without details.

Fig. 1: *Graph-Based Technique for Extracting Keyphrases*




978-1-5386-5697-6/18/$31.00 ©2018 IEEE
DOI 10.1109/ICPET.2018.00023

GTEK is based on three main advantageous factors, which are:

- Graph representation which is more accurate than the vector space model.
- Clustering the most similar sentences to cover the most important keyphrases.
- Use of the TextRank algorithm to score and rank the sentences.

The rest of this paper is organized as follows. We first briefly review the related work in the next section (Section II), followed by a description of document representation in our technique (Section III). Next, in Section IV, we present and explain our algorithm to extract the keyphrases. Finally, we discuss our experiments and the results in Section V and give our conclusions in Section VI.

## II. RELATED WORK

There are several classical methods to extract keyphrases and/or keywords. One of the easiest method is to consider the most frequent phrase or word in the keywords group. This method, however, generally yields bad results, because it does not consider many keyphrases if they have low frequency. Moreover, this method does not link between the importance of sentences and the selected keyphrases and word.

Recently, graph-based keyphrase extraction has received clear attention and many different methods and techniques have been proposed. All of these methods represent the documents using a graph structure in which nodes are words or phrases and edges represent co-occurrence or semantic relations [6]. The importance of each node is computed using a scoring algorithm such as TextRank [7], PageRank [8] or HITS [9]. Words corresponding to the top-ranked nodes are then selected and collected to generate keyphrases.

Google's PageRank algorithm [8] is one of the most popular link analysis algorithms and is used for web page ranking. It calculates the importance of a sentence based on the information extracted from the graph structure. The sentences that have a large PageRank value are important to be included in the final summary. The PageRank value is calculated based on the importance of the neighboring nodes and then is used for extraction of keyword/sentence.

In TextRank algorithm [7], graph-based model for ranking keyphrases and sentences. It depends on an enhancement algorithm derived from PageRank method. The text is modeled as either directed or undirected graph and as a weighted co-occurrence network. It is completely unsupervised and the extraction works as follows: First, build a word graph on a given document in which the links between terms describe their semantic connections, often are computed by the word co-occurrences in the document. Second, the score of each term is used to get the ranking candidate keyword by applying the PageRank algorithm on the graph. The authors of this work reported that their method beats classical statistics-based approaches with respect to precision and F-measure although the recall is lower than in the statistics-based methods. In TextRank, a low-frequency word will get the high score due to the existence of neighbors with high frequency which cannot occur in TFIDF.

In Hyperlink-Induced Topic Search (HITS) algorithm [10], calculation of the importance of a vertex is rely on the assumption that any term points to many others as (Hub) and pointed to by many others as (Authority). Therefore, this algorithm determines the values of authority and hub for any page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

An unsupervised graph-based keyword extraction method named Keyword Extraction using Collective Node Weight (KECNW) [17] was proposed to extract the keywords from the Twitter social networking site. It determines the weight of a keyword by taking various cumulative impact arguments such as frequency, position, and strength of its neighbors. The model was validated with five datasets, among them were Uri Attack and Harry Potter. The experimental results of that method were far better than the others in terms of precision, recall and F-measure. On the other hand, KECMW loses some advantages such as ignoring phrases with some importance and ignoring the relationship between words with the use of TF–IDF for calculating the similarity. Finally, the extracted keywords do not cover all topics of the tweets.

Finally, in [18], unsupervised methods based on the graph are used to extract the keyphrases and/or keywords. The principle of this method is to create a graph of words and/or phrases. The edges and their weights can be computed using co-occurrence counts or semantic relatedness using and without word embedding representation. Nodes are then ranked using graph properties such as centrality. The experiments show that word embedding representation does not improve the results significantly and it neither decreases the results.

As mentioned in the above section, the obtained keyphrases of all these methods are improper to cover all the topics in the document. To solve this problem, [11] proposed a clustering algorithm on the graph to detect the topics. The method leverages clustering techniques on the graph to obtain several clusters and then choose the candidate keyphrase close to the centroid of each cluster. Experimental results show that this approach performed better than the state-of-the-art keyphrase extraction method on two datasets under three evaluation metrics (precision, recall and F-measure).

## III. DOCUMENT REPRESENTATION

First, some basic concepts from graph theory are defined for describing the graph-based model then the Document Index Graph (DIG) model is described.

### *A. Graph principles*

A graph is an ordered pair of a set of vertices and a set of edges, in which each vertex represents an object of text as word and sentence, while each edge represents the

relationship between these objects. A graph is directed if the edges have a direction from one vertex to another and it is weighted if there is a weight function WF that calculates and sets a real value number to each edge.

Edge describes the relation between two terms which can be created on many principles based on different text scope or relations for the graph construction [14]:

1) Words arise as ordered pair in text.
2) Document added to the graph as a one object of group in ordering task .
3) Intersecting words from a text.
Additional description about graph structure and its measures can be viewed in [12] and [13].

### B. *The Document Index Graph (DIG) model*

GTEK exploits the Document Index Graph (DIG) model [15] in the stage of document representation which is directed graph and unweighted.

When coming up with a new document, DIG splits up the document into sentences and words. The obtained words are single characters different than phrases. Stop words will be removed from the words. Each word represents a vertex in the graph and each sentence is represented by a path between many words. Figure 2 describes the detailed data in the vertices of the DIG. Each vertex in the graph records all information about the word's occurrence in the sentences which makes the extraction of sentences from the graph simply available.

In DIG, the detection process of shared phrases is done while the graph is building, it means that when the graph is constructed, all the shared phrases among sentences of the document are available.

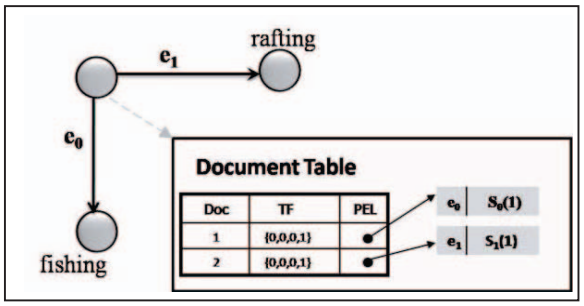
As shown above, GTEK does not use the graph model to rank the keyphrases but to represents the document as a connected group of terms. It then uses the clustering algorithm and TF to select the keyphrases as will be explained in the next section.

## IV. KEYPHRASES EXTRACTION

Extracting keyphrases from a single document in GTEK can be divided into four stages:

- *First*, Document Representation: represent the document as a group of vertices based on a graph as discussed in the previous section.

Fig. 2: *The structure of the data stored about each vertex of the DIG*




- *Second*, *Sentence Clustering*: discover the main topics of the document using clustering algorithm. (will be discussed in subsection A of this section)
- *Third*, *Sentence Ranking*: based on GTEK's motivation which considers the effect of the sentence's important on selected keyphrases, the sentences will be ranked using TextRank algorithm to detect the most important sentences. (will be discussed in subsection B of this section)
- *Fourth*, *Keyphrases Selection*: keyphrase candidates of each cluster are generated, then the most frequent one in the document will be selected. (will be discussed in subsection C of this section)

The following sub-sections describe the last three steps in detail.

### A. *Sentence Clustering*

The clustering algorithm of sentence is the core of the GTEK technique. Here, the widely used clustering algorithm Graph-based Growing Self-Organizing Map (G-GSOM) [15] is exploited to group the sentences of the document based on the graph model representation in order to discover the topics of document, as shown in Algorithm (1).

**Algorithm 1: Sentence Clustering Algorithm**

**I**nput: Document, Clusters Number (CNo), Similarity Threshold (ST)
**O**utput: Clusters of Sentences
Set counter equal *one*
Initialize the first cluster with the first sentence, then:
Set i equal *two*
While counter < size of document
  Calculate the distance between ($sentence_i$)and the clusters
  Find the most similar cluster (*similar)*
    if (distance > ST)
      Create a new cluster *newCluster*
      add $sentence_i$ in the *newCluster*
    else
      add $sentence_i$ in the *similar*
    end if
  Increment i by *one*
end while
Increment i by *one*

Measuring the distance between two sentences is based on the shared phrases and words between them. This means as shared phrases and words increase, the similarity will be increased. The distance between the two sentences using shared phrases is calculated by equation (1). The second part of similarity which depends on shared words is calculated based on the Euclidian distance similarity.

$$Sim_p(S_i, S_j) = \frac{\sum_{i=0}^{i=n} [(F_{1i} \times w_{1i}) + (F_{2i} \times w_{2i})] * Plenght_i}{(|S_1| \times w_1) + (|S_2| \times w_2)} \quad (1)$$

Where:

$F_{1i}$, $F_{2i}$: The frequency of the word *i* in sentence 1 and sentence 2, respectively with unique position.

$w_{1i}$, $w_{2i}$: The weight of the word *i* in sentence 1 and sentence 2, respectively.

$|S_1|$, $|S_2|$: The length of the sentence 1 and sentence 2, respectively.

$w_1$, $w_2$: The weight of sentence 1 and sentence 2 respectively.

$Plength_i$: The length of the $i^{th}$ shared phrase.

The number of clusters is determined by the length of the document and the distance between sentences. The preferred value will be explored through experiments. While assigning the sentences to the clusters, each sentence will be added as a new vertex to a sub-graph. This means that each cluster is represented as a graph of very similar sentences.

The clustering stage when finished produces several clusters each contains a set of sentences which are related to a certain topic. This cluster is represented by one graph in which each sentence is represented by a vertex. With this representation, the sentences are now ready for TextRank algorithm to be applied in the ranking stage.

### B. Sentence Ranking

In this stage, TextRank algorithm is applied on each cluster to rank the sentences in order to detect the most important sentences. After performing TexRank, each sentence in each cluster will be ranked according to its votes from others. Next, the most important sentence from each cluster will be selected to extract the candidate keyphrases; i.e, keyphrases in GTEK are extracted from the most important sentences in each sub-topic.

The number of selected sentences (*x*) is selected with consideration to the number of sentences in each cluster. The preferred value of *x* will be explored through experiments. In our experiments, *x* was set as discrete value from 1 to 5 based on the number of sentences in each cluster and based on the dataset.

### C. Keyphrases Selection

In this stage, all the phrases and words of the highest ranked sentences in the clusters will be considered as candidate keyphrases. These candidates will then be ranked based on their term frequency (TF) in the document. Here, redundant candidates will be filtered out. Two candidates are considered redundant if they have the same stemmed form (e.g. "argue", "arguing", and "argus" are reduced to the stem "argu") and if one phrase is part of the other.

GTEK gives a new value by its ability to extract the keyphrases that cover all sub-topics of the document and its consideration of the effect of sentence's important on extracted keyphrases.

## V. EXPERIMENTAL RESULTS

All experiments were carried out on two standard datasets in order to evaluate the performance of GTEK. The main one was the benchmark collection of news articles provided by the UCST [16]. The second dataset was built by Hulth2003 [19].

The UCST collection contained (330) English text articles, (106) news articles, (85) announcement articles, (98) programs' description articles and (41) event description articles. Each article had (3-8) golden standard keyphrases with average (4) keyphrases in each article. The number of keyphrases extracted in GTEK and by the expert based on the number of main sub-topics in the article. The Hulth2003 dataset, on the other hand, contained (1,460) abstracts of research articles. Each abstract had two kinds of manually labeled keyphrases.

In GTEK, a selected phrase or word is considered as keyphrase if it belongs to the highest ranked sentence in its cluster and it is one of the most frequents in the document. We create the keyphrases list for each document and compare it with the golden standard keyphrases.

For the purpose of evaluation, the commonly metrics are used, which are: precision, recall and F-measure. Equations (2) to (4) below show their definitions.

$$\text{precision} = \frac{c_{correct}}{c_{exctracted}} \tag{2}$$

$$\text{recall} = \frac{c_{correct}}{c_{standard}} \tag{3}$$

$$F - measure = \frac{2.p.r}{p + r} \tag{4}$$

In Table 1 and Table 2, GTEK is compared with two baselines, TextRank and TF-IDF. The results show that both methods fail to consider the effect of sentences on the importance of the word. Moreover, they fail to cover the sub-topics of the document. Overly, we observe that GTEK gives results better than TextRank and TF-IDF.

Table 1 presents the performance of GTEK compared with the other two algorithms performed on the UCST-news articles dataset while Table 2 presents GTEK results on the Hulth2003 dataset.

It is observed from the GTEK's results that each article has keyphrases from the most sub-topics of the article and from the most important sentences. This implies that GTEK is more capable to retrieve the most relevant instances among the retrieved instances.

Figure 3 presents a part of the sample article number (416) from the collection and Figure 4 presents the golden keyphrases of the article (416). The extracted keyphrases for the same article are shown in Figure 5.

TABLE 1: RESULTS ON UCST-NEWS DATASET

| Method | Recall | Precision | F-measure | Covering |
|---|---|---|---|---|
| GTEK | 76.7 | 86.8 | 81.1 | 85.2% |
| Text Rank | 48.6 | 50.0 | 49.2 | 44.3% |
| TF - IDF | 34 | 33 | 33.5 | 32.7% |

TABLE 2: RESULTS ON HULTH2003

| Method | Recall | Precision | F-measure | Covering |
|---|---|---|---|---|
| GTEK | 75.2 | 82.3 | 78.6 | 87.3% |
| Text Rank | 40 | 41 | 40.5 | 48.2% |
| TF - IDF | 32 | 31 | 31.5 | 34.9% |

*Fig. 3: Text document from UCST 2018 collection, part from article 416*

**Dr. Mohammad Sadiq has received his task as a dean college for the university college of science and technology.**

The College of Education Former dean college within Al-Aqsa University Dr. Mohammad Sadiq has received his new position as a Dean College for the university college of science and technology succeeding for Dr. Ziad Thabet who is returned to the Ministry of Higher Education as Assistant Undersecretary for Educational Affairs.

Dr. Mohammad Sadiq is considered as one of the most expert educators. He is held the Doctoral degree in educational psychology and he has many researches as well as he was participated in a number of educational conferences; the last year, he headed the International Conference of Education College. He was held several positions, latest of which was a dean college for Al-Aqsa University, a Secret keeper and a member of the National Committee for

*Fig. 4: Golden standard keyphrases for article 416*

dean college, university college, science and technology, new position, Mohammad Sadiq

*Fig. 5: Extracted keyphrases for article 416*

dean college, university college , higher education, new position, mohammad sadiq

The number of clusters (sub-topics) is not predefined. It depends on the content of the document. It is to be noted that many of the clusters were skipped when the cluster contained one sentence as its importance was less than the others. Overly, the number of clusters in the Hulth2003 dataset was either 2 or 3 because the document is composed of short abstracts which have less words. On the other hand, the number of clusters in the UCST dataset was 3 to 5 because the UCST dataset is composed of a full article about an event in an academic institution.

Values of precision, recall, and F-measure for (20) sample articles as dataset are graphically shown in Figure 6.

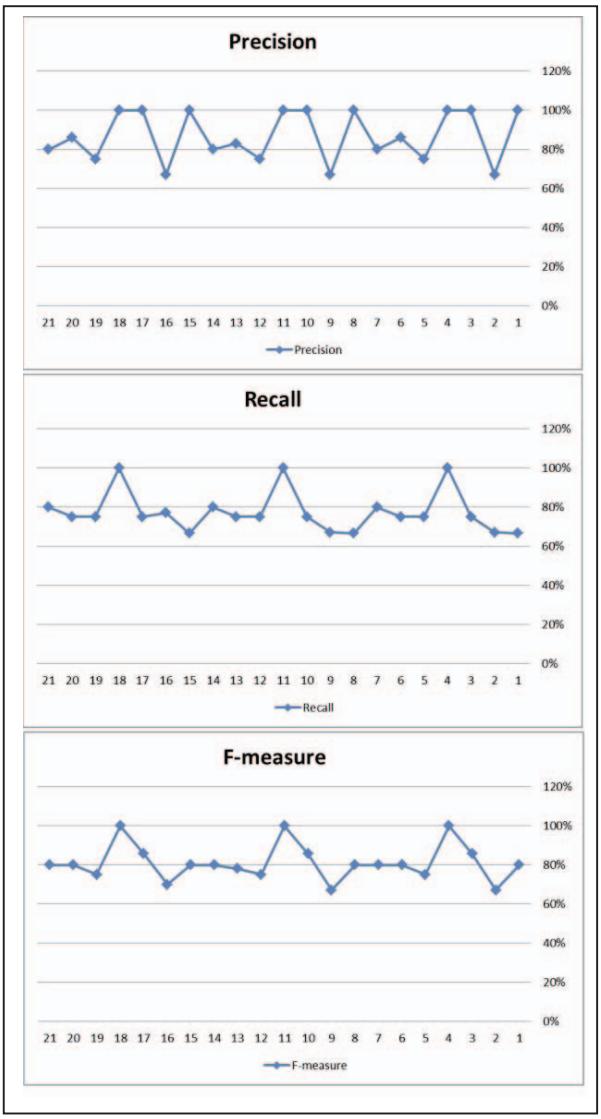
## VI. Conclustions

We have proposed GTEK as a graph-based representation model incorporating with GHSOM as clustering algorithm for clustering sentences to extract keyphrases of a single document. The extracted keyphrases cover the most important sentences and the main sub-topics in a document. Experiments showed that GTEK performs better than other baseline methods on two datasets. GTEK represents the document using the graph model and applies the GHSOM algorithm to cluster the sentences and then rank them using TextRank algorithm. GTEK produces improved results compared with TextRank and TF-IDF on two datasets.


## Acknowledgments

The authors grateful Prof. Ahmed Salahedden for his helpful comments on this work.


*Fig. 6: Values of evaluation results for samples of datasets*




## References


[1] M. Grineva, M. Grinev, Lizorkin. Extracting Key Terms from Noisy and Multitheme Documents. In Proceedings of the 18th International Conference on World Wide Web, 2009, Pages 661-670, NY, USA.

[2] M. Berry, J. Kogan, Text Mining: Applications and Theory, Wiley, UK, 2010.

[3] J. Manyika, J. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Byers, Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011.

[4] Y. Yan, Q. Tan, Q. Xie, P. Zeng, P. Li, A Graph-based Approach of Automatic Keyphrase Extraction, Procedia Computer Science,Volume 107, 2017, Pages 248-255, ISSN 1877-0509.

[5] S. Beliga, A. Meštrović, S. Martinčić-Ipšić, An Overview of Graph-Based Keyword Extraction Methods and Approaches. Journal of Information and Organizational Sciences, Volume 39 (1), 2015, Pages 1-20.

[6] B. Florian, A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction, IJCNLP, Oct 2013 Nagoya, Japan. pp. 834-838, 2013.

[7] R. Mihalcea and P. Tarau, TextRank: Bringing order into texts. In Proceedings of EMNLP, 2004.

[8] B. Sergey and P. Lawrence, The anatomy of a large-scale hypertextual Web search engine. Computer Networks, Volume 30(1–7), 1998, Pages 107–117.

[9] M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, Volume 46(5), 1999, Pages 604–632.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, Volume 46(5), 1999, Pages 604–632.

[11] Y. Ying, T. Qingping, X. Qinzheng, Z. Ping, and L. Panpan, A Graph-based Approach of Automatic Keyphrase Extraction. *Procedia Comput. Sci, Volume* 107,2017, Pages 248-255.

[12] J. Borge-Holthoefer, A. Arenas, Semantic Networks: Structure and Dynamics, Entropy, Volume 12(5), 2010, Pages 1264-1302.

[13] M. Newman, Networks: An Introduction, Oxford University Press, 2010.

[14] S. Sonawane, P. Kulkarni, "Graph based Representation and Analysis of Text Document: A Survey of Techniques", in Int. Jour. Of Computer Applications, volume 96(19), 2014, Pages 1-8.

[15] M. Hussin, M. Farra, and Y. El-Sonbaty, "Extending the Growing Hierarchal SOM for clustering documents in graphs domain," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, Pages 4028-4035.

[16] News articles provided by the University College of Science and Technology (UCST), 2018, available at www.cst.ps/en

[17] S. Biswas, M. Bordoloi and J. Shreya. A Graph Based Keyword Extraction Model using Collective Node Weight. Expert Systems with Applications. 2017, Pages 51–59.

[18] J. Mothe, F. Ramiandrisoa and M. Rasolomanana. Automatic keyphrase extraction using graph-based methods. 2018. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (SAC '18). ACM, New York, NY, USA. Pages 728-730.

[19] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. 2003. In Proceedings of EMNLP. Pages 216–223.