© Journal of the Technical University - Sofia
Plovdiv branch, Bulgaria
"Fundamental Sciences and Applications"
Vol. 26, 2020

A Review of Approaches for Keyphrase Extraction

Özlem Aydın, Fatma Büyüksaraçoğlu Sakallı, Andaç Şahin Mesut

Trakya University, Ahmet Karadeniz Campus, Department of Computer Engineering Edirne/Turkey

ozlema@trakya.edu.tr, fbuyuksaracoglu@trakya.edu.tr, andacs@trakya.edu.tr

Abstract. In this paper, a survey of approaches for Keyphrase Extraction (KE) is presented. KE is an important task for Text Mining (TM), Information Retrieval (IR) and Natural Language Processing (NLP). The main aim of KE is to extract the most important words in a text. This article introduces a survey of approaches for KE task. In addition to analysing KE approaches, KE studies for the Turkish language were also evaluated in a different section.

1. Introduction

Keywords/keyphrases are the smallest units that can summarize the content of a text. A keyword is a single word (e.g. Internet, data), whereas a keyphrase is a combination of keywords which makes phrases (e.g. data science, Natural Language Processing). They are used for identify information most relevant of the content of a text. For example, one of the places where keywords/keyphrases are used are academic publications. The words/phrases that best represent working in such publications are mentioned at the keyword section at the beginning of the paper. The reader can easily decide whether to read the full text or not through keywords/keyphrases.

In the digital environment, text-based data is huge and growing very fast. When searching any subject on Internet, there may be difficulties in accessing the subject-related text. If the search at big data takes place over the topics related to the searched, access to the requested information will be correct. If there is a subset of words (keywords) that represent the main features, content, theme, and similar features of the texts, it can be made easier to analyse such large amounts of data. Keyphrases can be considered as the smallest summary of a text and can serve to easily organize the texts and reach them according to their content.

Keyphrase Extraction (KE) is one of the important task in many fields. [1] present some tasks in Text Mining (TM), Information Retrieval (IR) and Natural Language Processing (NLP) which are widely used for KE. These tasks listed below:

- Document Clustering
- Document Summarization
- IR Systems
- Document Indexing
- Web Mining
- Search Engines
- Query Refinement
- Web Logs
- Recommender Systems
- Opinion Mining
- Relevance Feedback
- Ontology
- Information Extraction
- Named Entity Recognition
- Topic Analysis

Our concern in the work presented in this paper is a survey of approaches for KE and some KE studies for Turkish language. In the rest of this paper, we first summarize some previous related studies for KE approaches which are supervised, unsupervised and deep learning (cf. Section 2). Afterwards,

 Online ISSN 2603-459X

we introduce KE studies for the Turkish language (cf. Section 3). Finally, we offer a brief conclusion including suggestions about KE studies (cf. Section 4).

2. Keyphrase Extraction Approaches

KE is realized with three different approaches: supervised, unsupervised and deep learning. In the studies conducted to date, the methods applied with unsupervised approaches have become more popular. Because they are domain independent and do not need labelled training data. On the other hand, supervised methods have more powerful modelling abilities and ordinarily achieve higher accuracy than supervised methods [2].

2.1. Supervised Approaches

KE task was considered as a binary classification problem in early supervised approaches. In this approach, a classifier is trained on documents annotated with keyphrases in order to determine whether a candidate phrase is a keyphrase or not.

The features generally used to represent an instance for supervised KE can be divided into two categories [3]: within-collection features, external resource-based features. Within-collection features are computed based solely on the training documents. These features can be further divided into three types: statistical features, structural features and syntactic features. *Statistical features* are computed based on statistical information gathered from the training documents. Three such features have been extensively used in supervised approaches: TF-IDF, the distance of a phrase and supervised keyphraseness. *Structural features* encode how different instances of a candidate keyphrase are located in different parts of a document. A phrase is more likely to be a keyphrase if it appears in the abstract or introduction of a paper or in the metadata. *Syntactic features* encode the syntactic patterns of a candidate keyphrase. For example, a candidate keyphrase has been encoded as (1) a POS tag sequence, which denotes the sequence of part-of-speech tag(s) assigned to its word(s); and (2) a suffix sequence, which is the sequence of morphological suffixes of its words.

External resource-based features are computed based on information gathered from resources other than the training documents, such as lexical knowledge bases (e.g., Wikipedia) or the Web, with the goal of improving KE performance by exploiting external knowledge.

KEA [4] and GenEx [5] systems are the most popular systems developed by supervised methods. In these systems, the frequency and location of the candidate keyhrase in the document are the most important features used for classification. KEA calculates the TF-IDF value and the first occurrence for each candidate keyphrase. Naive Bayes method is used in training and keyword extraction stages in KEA system. GenEx is based on a genetic algorithm that optimizes the number of correctly defined keywords in training documents. This algorithm consists of Genitor genetic algorithm and Extractor KE algorithm. C4.5 decision tree was used for learning for GenEx system.

Unlike the KEA and GenEx models, there is no limit for the length of keyphrases in the Hulth model [6] which was developed with another learning based method called bagging. While applying bagging technique, four different features were used: term frequency, collection frequency, the relative position of the first occurrence of a term and POS tag term.

A KE system called KPSpotter which is information gain-based was developed by [7]. [8] propose a new method called KEA++, which enhances automatic KE by using semantic information on terms and phrases gathered from a domain-specific thesaurus.

In the [9] study, which is classified with Decision Support Machines, keyphrases are grouped in three different classes: good keyword, indifferent keyword and bad keyword. According to the results obtained, this proposed SVM-based method performed significantly better than basic methods for KE.

[10] have identified three different features in the HUMB system they developed: structural features, content features, lexical/semantic features. Decision Tree (C4.5), Multi-layer Perceptron and Support Vector Machines are used together in the ranking section of their system.

2.2. Unsupervised Approaches

Unsupervised approaches have been examined in four different groups listed under.

2.2.1 Graph-Based Ranking

KE is a task in NLP, where the goal is to identify the most important words and phrases in a document. The importance of a candidate keyphrase is defined by how much it relates to other candidates keyphrases in the document. If a candidate keyphrase is related to a large number of candidate keyphrases and these candidate keyphrases are important, they will also be important. A graphic is created from the input text and its nodes are ranked by a graph-based ranking method according to their severity. Each node in the graph corresponds to a keyphrase, while the edges connect the two candidate keyphrase. Edge weight gives the proportion of the syntactic and/or semantic relationship between connected candidates. For each node, each of its edges is evaluated by a vote from other nodes to which it connects with the edge. The core of a node in the graph is defined recursively by the edges it has and the scores of neighboring nodes. The candidates at the top of the ranking in the graph are selected as keyphrases of the input text [3].

TextRank [11] is one of the most known graphic-based ranking methods for KE. In this method, the relationship between candidate keyphrases is determined using co-occurence numbers.

SingleRank [12] which is a variation of TextRank that incorporates weights to edges. In this method, unlike TextRank, the number of occurrences between two words is important and this value is used to calculate the weight of the edges. The other difference is that SingleRank does not filter out any low scored words.

CiteTextRank [13] used information from citation networks for KE. The KE task was carried out on research articles. CiteTextRank uses document content and other contents in the citation network to which the document refers.

In the RAKE [14] algorithm, the stop words are extracted from the words in the document and the remaining words are determined as candidate keywords. Then, a score is calculated for each candidate keyword. RAKE utilizes word frequency and word degrees to assign scores to keywords.

2.2.2 Topic-Based Clustering

In this approach, candidate key phrases are grouped according to their subjects. The perspectives brought by the approach are [3]:

- A keyphrase relates to one or more topics in a document.
- The extracted keyphrases should be inclusive of all topics in a document.

In topic-based methods, clustering techniques and Latent Dirichlet Allocation are used to identify main topics.

KeyCluster [15] uses Wikipedia and co-occurence-based statistics to group semantically similar candidates. The other method called TopicRank [16] is an improved variation of TextRank. The noun phrases (NPs) that determine the main topics in a document are selected, then these NPs are grouped according to their topics and are determined as edges in a full graph. Later, TextRank is applied to determine the score of the topics and the keyphrase extraction is completed by selecting the best candidate keyphrase representing each top-ranked topic.

2.2.3 Simultaneous Learning

It is a method that arises with the assumption that text summarization and KE will benefit each other if they are performed simultaneously. [17] proposed the first graphic-based approach to realize with this assumption. The idea in this approach is that if a sentence contains important words, it is important and important words are found in important sentences.

[18] expanded Zha's work by adding two assumptions: an important sentence depends on other important sentences, and an important word is linked to other important words [3].

2.2.4 Language Modelling

[19] propose an approach that combines steps which are extracting candidate keyphrases and ranking keyphrases. They applied two language models for scoring the phraseness and informativeness of phrases. Phraseness feature is identified as extent which a sequence of multi word can be treated as a phrase. The informativeness feature determines whether a phrase provides information about the document.

2.3 Deep Learning Approaches

Deep learning is a subfield of machine learning based on learning data representations. Below is a definition for deep learning by [20]: "Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.". The use of deep learning in NLP, a field that needs high computing and data processing, has brought successful results.

[21] proposed a Recurrent Neural Network (RNN) model to extract keyphrases from tweets. This RNN model is a combination of keyphrases and content information. The target layer is defined by combining the two output layers. In the model, which has two hidden layers, the last hidden layer and the outputs from the previous hidden layer are combined with the linear regression function and form the target layer.

In another study, the Convolutional Neural Network (CNN)-based method called CopyCNN was proposed [22]. In this method, after the candidate keywords in the text are removed, the copy mechanism is used. In order to obtain more successful results, importance mechanism and location information have been added to the model.

A RNN-based generative model was designed by [23] to predict keyphrases. An RNN-based producer model was designed to predict keyphrases. In the model they designed, CopyRNN, they added a copy mechanism to RNN that allows them to successfully guess the "secret keyphrases" that are rare. This mechanism used the location information of words when calculating the importance vector of the word.

[24] presented the Title Guided Network (TG-Net) for KE as a new model. This model has the encoder decoder architecture, which has two new features: (1) the title information is also used as a query-like input, and (2) an encoder with a title guide collects relevant information from the title for each word in the document. A separate two-way Gated Recurrent Unit (GRU) encoder is defined for the title and body text in TG-Net.

3. Some Turkish KE Studies

There are very few KE studies conducted for the Turkish language to date. The first study in this field was carried out by [25]. They applied the KEA algorithm, which has a place in the literature, for the Turkish language. In implementation, they replaced the original stemmer of this algorithm and the list of stopwords with their Turkish counterparts. In addition, they incorporated the *relative length* feature as a new feature not found in the KEA algorithm.

In another study [26], noun phrase, noun phrase (NP heads), length and first occurrence statistical data were used. It is similar to the B&C method in that it does not require corpus training. On the other hand, it also uses some features calculated with KEA and GenEx methods. It showed a similar performance with a study on a similar review [25]

[27] applied the Multi-Criterion Ranking (MCR) method for KE. Their method consists of two stages. In the first stage, candidate keyphrases are extracted from the text and calculated their scores wit features. In the second stage, a Hasse diagram is created from candidate keyphrases. Then keyphrases that are suitable with MCR are selected. Corpus training is not required. More successful results were obtained from the TurKeyX [27], and it showed close success with KEA-TR [25].

[28] conducted KE using the academic articles. They used prepositions and conjunctions from the texts by using NLP methods. Afterwards, they calculated TF-IDF values and used the TextRank algorithm to determine keyphrases.

Messages accumulated within the 7/24 Yıldız Line Management System, which is actively used by internal stakeholders at Yıldız Technical University, were used for keyword feature extraction [29]. Significant and useful keywords were found in the results of analysis conducted with ChiKare, Information Gain and TF-IDF methods.

Turkish Labeled Text Corpus [30] is prepared for the need for in Turkish KE studies. A labeled text corpus consist of Turkish papers' titles, abstracts and keywords. Although it is a compilation created especially for text classification studies, it can also be used in some areas such as KE, title extraction and text summarization due to its content.

4. Conclusion

In this study, the studies in the literature for KE were examined. The survey work was grouped under three categories such as supervised, unsupervised and deep learning approaches for KE and some important works done in those areas were listed chronologically. Since deep learning has yielded better results than the existing methods in solving many problems in recent years, it has been also applied in the KE field. However, there are only a few KE studies in this field because of fact that training deep learning architectures require a large amount data which are expensive to acquire. When these constraints are eliminated, it can be said that the performance with deep learning methods will be better than other approaches in the KE studies. In addition, some studies on the Turkish language up to today are examined in a different section. It is seen that the studies for Turkish are very few. If large datasets are created for the Turkish language and more successful NLP tools are developed, it is thought that the studies in this area will increase.

References

[1] Merrouni Z A, Frikh B and Ouhbi B 2019 *Automatic keyphrase extraction: a survey and trends* Computer Science Journal of Intelligent Information Systems
[2] Papagiannopoulou E and Tsoumakas G 2019 *A review of keyphrase extraction* CoRR, abs/1905.05044
[3] Hasan K S and Ng V 2014 *Automatic keyphrase extraction: A survey of the state of the art* In Proceedings of ACL
[4] Witten I H, Paynter G W, Frank E, Gutwin C and NevillManning C G 1999 *Kea: Practical Automatic Keyphrase Extraction* In Proceedings of the 4th ACM Conf. of the Digital Libraries, Berkeley, CA, USA
[5] Turney P D 2000 *Learning algorithms for keyphrase extraction* Information Retrieval
[6] Hulth A 2003 *Improved automatic keyword extraction given more linguistic knowledge* EMNLP pp 216-223
[7] Song M, Song I-Y and Hu X 2003 *KPSpotter: a flexible information gain-based keyphrase extraction system* In Proceedings of 5th Int. Workshop of WIDM 2003 pp 50-53
[8] Medelyan O and Witten I H 2006 *Thesaurus Based Automatic Keyphrase Indexing* In Proceedings of the 6th ACM/IEEE-CS JCDL pp 296-297
[9] Zhang K, Xu H, Tang J and Li J Z 2006 *Keyword Extraction Using Support Vector Machine* In: Proceedings of the Seventh International Conference on Web-Age Information Management (WAIM2006) Hong Kong China pp 85-96
[10] Lopez P and Romary L 2010 *HUMB: Automatic key term extraction from scientific articles in GROBID* In Proceedings of the 5th International Workshop on Semantic Evaluation pp 248-251
[11] Mihalcea R and Tarau P 2004 *TextRank: Bringing order into text* In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing
[12] Wan X and Xiao J 2008 *Single document keyphrase extraction using neighborhood knowledge* In Proceedings of the 23rd AAAI Conference on Artificial Intelligence pp 855-860
[13] Gollapalli S D and Caragea C 2014 *Extracting keyphrases from research papers using citation networks* In Proceedings of the 28th AAAI Conference on Artificial Intelligence pp 1629-1635
[14] Rose S, Engel D, Cramer N and Cowley W 2010 *Automatic keyword extraction from individual documents* Text Mining:Applications and Theory pp 1-20
[15] Liu Z, Li P, Zheng Y and Sun M 2009 *Clustering to find exemplar terms for keyphrase extraction* In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing
[16] Bougouin A, Boudin F and Daille B 2013 *TopicRank: Graph-based topic ranking for keyphrase extraction* In Proceedings of the 6th International Joint Conference on Natural Language Processing pp 543-551
[17] Zha H 2002 *Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering* In Proceedings of 25th Annual International ACM SIGIR

Conference on Research and Development in Information Retrieval pp 113-120.
[18] Wan X Yang J and Xiao J 2007 *Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction* In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics pp 552-559
[19] Tomokiyo T and Hurst M 2003 *A language model approach to keyphrase extraction* In Proceedings of the ACL 2003Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18. Association for Computational Linguistics MWE'03 pp 33-40
[20] LeCun Y, Bengio Y and Hinton G. 2015 *Deep learning*. Nature 521 pp 436-444
[21] Zhang Q, Wang Y and Gong Y 2016 *Keyphrase extraction using deep recurrent neural networks on twitter* In Proceedings of the 2016 conference on empirical methods in natural language processing
[22] Zhang Y, Yang F and Xiao W 2017 *Deep keyphrase generation with a convolutional sequence to sequence model* 4th International Conference on Systems and Informatics (ICSAI)
[23] Meng R, Zhao S, Han S, He D, Brusilovsky P and Chi Y 2017 *Deep keyphrase generation* In 55th ACL, volume 1 Association for Computational Linguistics pp 582-592
[24] Chen W, Gao Y, Zhang J, King I, and Lyu M R 2019 *Title-guided encoding for keyphrase generation* In Proceedings of AAAI Conference on Artificial Intelligence
[25] Pala N and Cicekli I 2007 *Turkish keyphrase extraction using kea* In Proceedings of the 22nd International Symposium on Computer and Information Sciences (ISCIS 2007) Ankara, Turkey pp 1-5
[26] Kalaycilar F and Cicekli I 2008 *TurKeyx: Turkish keyphrase extractor* in Proceedings of the 23rd International Symposium on Computer and Information Sciences (ISCIS 2008) Istanbul Turkey pp 1-4
[27] Ozdemir B and Cicekli I 2009 *Turkish Keyphrase Extraction Using Multi-Criterion Ranking* In: 24th International Symposium on Computer and Information Sciences pp 269-273
[28] Yıldız O 2017 *Metin Madenciliğinde Anahtar Kelime Seçimi Bir Üniversite Örneği* Yönetim Bilişim Sistemleri Dergisi 2(3) pp 29-50
[29] Müngen A A and Kaya M 2018 *Extracting abstract and keywords from context for academic articles* Soc. Netw. Anal. Min. vol.8 no.1 p 45
[30] Ozturk S, Sankur B, Gungor T and Yilmaz M B 2014 *Turkish Labeled Text Corpus* 22nd Signal Processing and Communications Applications Conference (SIU)